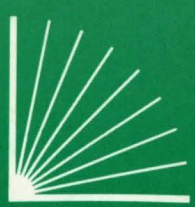
Rapport 20/91

Research Evaluation

Proceedings of a Conference

Oslo, 30–31 May 1991



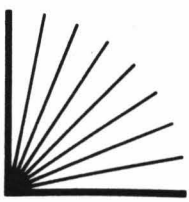
Utredninger om forskning og høyere utdanning
NAVFs utredningsinstitutt
Norges allmennvitenskapelige forskningsråd

Rapport 20/91

Research Evaluation

Proceedings of a Conference

Oslo, 30–31 May 1991



Utredninger om forskning og høyere utdanning
NAVFs utredningsinstitutt
Norges allmennvitenskapelige forskningsråd

ISBN 82-7218-272-6
ISSN 0802-9342

GCS AS - OSLO

Foreword

This report contains the proceedings of a conference on research evaluation held at Holmenkollen Park Hotel Rica, Oslo, 30-31 May 1991. The aim of the conference was to discuss how to perform research evaluations of high quality. The target group consisted mainly of staff in the research councils responsible for evaluative work.

The conference was arranged by the Joint Board of the Norwegian Research Councils and the Institute for Studies in Research and Higher Education, the Norwegian Research Council for Science and the Humanities. The programme committee consisted of Arne Berge, the Joint Board of the Norwegian Research Councils; Svein Kyvik, the Institute for Studies in Research and Higher Education; and Kirsten Voje, the Royal Norwegian Council for Scientific and Industrial Research.

Svein Kyvik and Sue Ellen Walters have edited the final report.

Oslo, December 1991

Johan-Kristian Tønder

Hans Skoie

Contents

James McCullough

Evaluation in the National Science Foundation

I am going to talk about the National Science Foundation in the US Government and its role vis-à-vis the other US Government agencies and other sources of research support in the United States, then about the Program Evaluation Staff which is a small unit in the Director's Office which I lead, and then of several specific evaluations which we have completed and some others which are underway.

The US National Science Foundation was started in 1950. It was by political decision based on a realization that science had contributed a lot to the US effort in the Second World War. But it was meant to be set up as a civilian agency for support of science. The military services have their own research support agencies, so we do not undertake to fund any classified research. For that matter we do not fund any medical research. That is the province of another organization; the National Institutes of Health, NIH, in the US.

We have been in business now for forty years and I think the Foundation has made just about all the mistakes that can possibly be made, although we find new ways to make them. But we have a very good reputation, I think it is fair to say, with both the President and the Congress. We are probably, however, not well known with the public, outside of the public of science. They confuse us often, for example, with our National Academy of Sciences and think that we are the same organization. We fund research, make grants for research, in all fields of science and engineering, including computer sciences, and social sciences. We have responsibility for making grants in areas for improving the teaching and learning of science in the schools. Now we can't do very much to influence education in the grammar schools and high schools in the United States. There are some seventeen thousand independent school systems across the US. But what we can do is help people develop better teaching materials, better films and science books and so forth. We also support a number of the popular science programs on television.

We also have responsibility for the US effort in the Antarctic. You may know that we have a very active base there at the south pole with quite a long logistical supply line. The rationale for US presence in the Antarctic is scientific rather than military and that is why even though a lot of the organization is under the US Navy, it is basically budgeted and programed in the National Science Foundation.

Most of our budget goes to support individual small research projects by professors at universities. But we also fund large operations like telescopes, for example, in the southwestern United States and research ships and several supercomputing centers to which scientists can take their various research problems and have them worked at very fast speeds on supercomputers.

NSF was founded to support the advance of research in various fields for its own sake, for the sake of science, in contrast to what are called in the United States the mission agencies. That is to say, for example, the Institutes of Health support research to advance the health mission and not so much for the advancement of research per se. The Department of Energy which has a component that came from a nuclear regulatory commission supports work on nuclear physics, in connection with its broader Energy mission. The Defense Department supports basic research in various areas to support its own Naval and Air Force missions, Army missions, weapons development and so forth. So that our role in funding various fields in part has to do with what other federal agencies are doing. For example, in computer science the defense agencies support a lot of work so we will support jointly perhaps some work in those areas, but they have the majority of the work. In biology the National Institutes of Health support most of the work, we have some role in funding molecular biology, cellular biology, outside the medical schools.

I was mentioning about the proportion of NSF funding in different fields. If you look at our budget you will see that in some areas, like mathematics, we supply a lot of money because we are the central source for funding of mathematicians, also of astronomers, let us say. However, in some areas, like biology, we are a secondary source to other organizations like the National Institutes of Health. We have a budget of two and a half billion US dollars, which is indeed a lot of money. But we have also twenty-eight thousand proposals a year. If you think about the scale of the university system in the United States, there are at the top perhaps fifty to one hundred top-notch research universities on the order of Stanford, Massachusetts Institute of Technology, Illinois and so forth, all of which have very strong departments in many areas. Each of those universities, big research universities, may send us two or three hundred proposals a year. And then we have a couple hundred other colleges and universities, each of which has maybe three or four strong departments, say in mathematics, in chemistry and biology. They may send us fifty or a hundred proposals a year. And then we have a number of colleges which are principally set up for teaching, but the teachers do some research and they may send us perhaps ten or twenty proposals a year. When you take the whole scope of all the programs we get these twenty-eight thousand proposals a year.

We have a staff of twelve hundred people of whom four hundred are scientists and engineers. The rest are clerical staff, administrative staff, and because we are an independent agency we also have to have organizations that deal with the

Congress, that deal with the President and the White House and so forth. So we have compressed into one small agency things that you would find in several layers in other bigger organizations.

We operate our program in units called divisions. All of our chemistry programs, for example, we may have six or eight different granting programs in chemistry, would be in one division. Mathematics has maybe another six or eight different programs. We have more than thirty-five divisions, including as well as engineering as I mentioned, social sciences and science education. So, we have all together about two hundred programs each of which has its own budget for making grants in a specialized field of research. And we have about three hundred program officers, so quite commonly a program will be operated by one program officer with a portfolio perhaps of a hundred or a hundred and fifty new applications coming in and perhaps another two or three hundred that have been granted and are being monitored at one time.

All proposals are reviewed externally. We require a minimum of three external reviews. There is an exception which I will tell you about in a minute. We require three external reviews before a decision can be made, commonly we have five or six reviews. Now, how this is done depends on the tradition in the program, in the particular field. About a third of the proposals are reviewed in what is called an *ad hoc* manner, that is, by the program officer.

We have I think you might say a strong program officer system where the program officer makes the decisions. Usually he or she has a doctorate, has done some research in the field, and is very knowledgeable about the field. I will contrast it with the National Institutes of Health, where the panels of reviewers are very strong and the program officer is called an executive secretary, relatively different. But in the NSF system the program officer is responsible for selecting the reviewers, for seeing that the reviews are competently done, and for integrating the results of the reviews and for making a recommendation about whether the proposal should be granted or not.

In the mathematical and physical sciences programs generally the custom is to do this through the mail, sending it out throughout the country to several reviewers and getting the results back with the program officer writing then a justification with a recommendation as to what should be done. In the biological and social sciences, the custom there is to send for a few reviews, and bring those to a panel and perhaps handle twenty-five or fifty proposals in two or three days with a sitting panel of eight or ten people who also have the benefit of having two or three reviews on each proposal from people other than the panel. In some areas where we have say equipment grant programs for colleges, for scientific teaching equipment, where proposals will be very very similar, (unlike the research programs

where the proposals' content will be quite different), we will have panels meet and just the panels make decisions.

Quite frequently when the panels meet they will group proposals according to very broad categories, perhaps the top twenty or twenty-five percent. They will say these are the best proposals, these should be funded. Then they will take the next twenty or twenty-five percent and say these are very good proposals, but the Foundation should pick and choose where it wants to fund them. And then certain criteria may come into play perhaps in some cases putting money around in different institutions or different parts of the country, or helping young people get a start, or helping women into careers in science, for example. So the proposals are not rank ordered very strictly. There are put into groups where the panels are advising the program director.

I contrast this with another style that is used in the US, that's by the National Institutes of Health, where there is a very strong panel system. All proposals go to the scientific review panels which very strictly rank each one and give them scores virtually to two or three decimal places. Now, I think this is an extremely hazardous kind of thing to do because I do not think that those fine distinctions can be made among proposals. But, never mind, what they do is rank them very strictly and then they fund from the top right down to when they run out of money. Then they draw the line and above that is called the payline and below that doesn't get funded. The NSF system is much looser and much more given to the judgement of the program officers.

Let me now mention my small evaluation staff, starting with its function and its place in the organization. We are part of the Director's Office. The Director's Office contains a number of small offices that you might find at the head of any agency, for example our legal counsel, congressional relations, budget and planning office and so forth. The evaluation function is a program evaluation function. We don't look at proposals. We try to look at how whole programs are operating. It is part of the budget and planning office, so it is linked with forward planning. And the idea is that forward plans are done, the budgets are written from the plans, the money is spent in accordance with the budgets, then we try to find out what we are getting for our money through our programs - which is linked back again with planning.

NSF has other types of evaluation functions, for example, we have an Inspector General's Office which is quite separate from my office and from the rest of NSF, and which is looking to see that proper procedures are practiced and that money is spent correctly and accounted for properly. That is not the function of my office; we try to be very constructive in terms of making recommendations to management as to how to improve programs. So our principal function is to do systematic

studies of programs and of our own proposal review system and to advise management on how to improve them.

I was brought in by the Director about five years ago to revitalize this area. We had some people doing very highly technical studies which were not linked with management's interests and what the Director wanted to see in terms of evaluation. Nor were they being published, they were more or less being held in-house. So I was brought in to help link our activities to the Director's agenda and also to see that the studies are accessible. We now publish them very widely so that people in our research community, in our management, in the Congress and so forth can see them.

We have several lines of business. One is to look at the proposal review system itself, and we spent a lot of work on this the last few years trying to understand our own process better and to make some improvements in it. Secondly, some work is done on the general value of investment in research and the general value of investment in NSF programs. Thirdly, we have looked at some particular programs, a very small number because we have a small staff and it is hard sometimes to choose which ones to look at. Fourthly, we have helped our own program managers, we have advised them on how to set up their own evaluations when they want to do their own work.

Now the Foundation, and my office, functions live in a context where a lot of evaluation is already going on. I mentioned that each program evaluates the proposals and then spends some time monitoring to see what is happening with those. Also in the US system you have large organizations, professional societies for example like the American Chemical Society, which every few years will issue a report about the priorities in the field to try to influence budgets and try to influence priorities. So, for example, about five years ago they came out with a report named after its chairman, the "Pimentel Report" in which they stated what kinds of new breakthroughs could be expected on the horizon, what kind of equipment would be needed, what would be the justification for training more graduate students in chemistry, and so forth. The National Academy of Sciences, for example, does its own reports of this sort. There is one by a person who is now the science advisor to the President, Mr. Bromley, who did one in physics about ten years ago which addressed, for example, what are the areas of physics that need development, what are the big new pieces of equipment to be operated, and so forth.

Also our divisions, again let me refer to our chemistry division and our mathematics divisions, as examples, each has an advisory committee of scientists from throughout the country which meets once or twice a year, not to look at proposals but just to discuss the situation generally in their fields and to advise on where they think the program directors should be putting their effort and putting their money.

So, our programs get quite a lot of advice one way or another, also of course from people who submit proposals. We fund altogether about 30 percent of the proposals. So we have in any given year about 70 percent unhappy people who have been declined and they are inclined to present their views too as to why their areas of science should have more money.

Now, an example of some of the things we have been doing. I mentioned that we spend a lot of time on our proposal review system in part because our Congress has heard complaints about it. Is it fair? Is it an old boys' network? Is it too conservative? In 1987 I undertook the first comprehensive survey of all our applicants in a given year. There were 14,000 people who applied for research grants in 1986. And in 1987 we sent them all a sixteen page questionnaire with forty questions some of which were open-ended. We got back ninety-five hundred responses, two-thirds return on our survey which shows you how much they really cared about all this. Many of them had written very extensive answers to the narrative questions. In fact, we got quite bogged down in having far too much data. It took a long time and a lot of struggle to write a good report.

We learned a lot about the communities we were dealing with, for example, we were able to sort out attitudes about whether people were satisfied or dissatisfied with how the proposal was handled by creating a matrix of six different types that had experience with NSF. For example, some people had put in one proposal only in five years and had gotten awarded; they were very happy with the system. Many people had mixed experience of various sorts. They had put in four or five proposals, perhaps had two awarded or one awarded, sometimes all five awarded. Some people had tried four or five or six times and had been refused every time. In any event, we were able to match their views about satisfaction and dissatisfaction with their experience with whether or not they had gotten grants.

The most important thing we found was that the applicant community was saying something different to us than we heard from the political community. In the political community we were hearing that NSF does not make grants to various parts of the country because we are biased against, for example, the mid-western United States. Or we are not making grants to small colleges because we are in favor of big research universities. Or we are not making grants to minorities, and so forth. But that was not the case when you ask applicants, and we asked them a couple of ways to just express themselves as to what the problems were. They didn't say very much about the problems I mentioned. They said two other things. One is they were disappointed in the reviews themselves that they were getting back from their fellow scientists, that many of them were too brief. They were not thorough enough critiques to be useful in preparing another proposal. And secondly, the widespread opinion was that the review system was too conservative. Both the reviewers and the program officers were not willing enough to take chances, to take

risks, to try enough new things. One almost had to have a perfect proposal and had to have too much of the work already done before it could be approved. Now this is in a context where even though the amount of money had been growing slightly, the actual funding rate had gone down. We were approving about forty percent of the proposals ten years ago and thirty-one percent last year.

Concerning the issue of conservatism and riskiness, our Engineering Directorate had tried a pilot program of quick turnaround grants without any external peer review, that is decisions made just by the program officers alone for small amounts of money for one year. These un-peer-reviewed grants could not be continued or renewed; that is, if one wanted to work more in that area one had to come back with a full-scale proposal that would be externally reviewed. In 1988 we did an evaluation of that program and found that both the program officers and the people who had gotten the grants believed very strongly that they were doing a lot of work that would not have been funded otherwise, that would not have survived critical peer review because it was very preliminary work, but was nevertheless useful work. So we recommended to the NSF Director that authority for this be extended across the whole Foundation. Now each of our program officers can spend up to five percent of their existing budget on these small grants without external review, on the program officer's judgement alone. These are very brief proposals and applicants contact the program officer first to see if these exploratory proposals will be entertained or whether they should submit a regular proposal.

In the first year about five hundred and forty applications came in, and we funded about half of them - which is a higher funding rate than our usual rate, but also because one has to have a lot of contact with the program director first. About half of the applications were from people who had no prior contact with NSF. So it seems likely that we are encouraging people who for one reason or another either don't trust the system or don't want to write a long proposal or don't think they'll get funded because the system is too conservative, but who are willing to do something in a very brief way.

We devised this program to overcome the conservatism of the system and to see that perhaps our programs will be able to take more risks with small amounts of money for people who want to do very preliminary work. One criterion is a quick turnaround on a proposal because some data are only available for a short time. This came in very handy during and right after the earthquake in California last year and also the hurricane, Hugo, that hit the east coast of the United States last year. A number of engineers, sociologists, and other researchers wanted small amounts of money to do follow-up quick studies on these natural phenomena and on the effect on society and we were able to make grants right away without any further review or permission.

We have also done some studies of a program called Research Experiences for Undergraduates where undergraduate students in their second or third year of

college spend a summer in a group of perhaps fifteen or twenty with a scientist or engineer doing hands-on research. One of the US's problems is that the number of people getting science doctorates in universities is going down considerably. Our evaluations show that program was successful in getting students interested fairly early in committing themselves to become researchers.

Another area that we have been working in is the role of National Science Foundation funding in the careers of various researchers who had won prizes in different areas (not only the famous Nobel Prize). In different fields of science and engineering one can pick out top prizes; we picked about fifty of those. We wrote to the prize winners and asked them various questions about the funding support at different points in their career, as a student, as a postdoctorate, as a beginning professor, and so forth, and found that NSF had quietly played a very strong role in critically funding them for their prize-winning research and also throughout their careers.

We are also doing some work which I hope will be wrapped up in the next couple of months to look at a program during the 1980s where we put in large grants to build up about twenty-five computer science departments. This is one of the few areas that had really been growing in the US during that decade and needed not just support for individual investigators but for larger amounts of money to buy equipment and to form up departments in these areas. And we're looking at what can be learned from that effort and what it accomplished. Again using surveys, bibliometric methods, interviews and so forth and contrasting the departments that were funded with those that were not funded - that sort of work.

Let me end by saying that we take the philosophy of cooperating with the various programs that are being evaluated. We don't want to get into too much of an antagonistic mode. We work out our project designs with them and we circulate our draft reports to them for corrections for factual accuracy and so forth. But we are very aware of the need to be autonomous, to be independent and to have our own voice. And so far we have not had a problem with having our reports suppressed or trying to have them changed too much by the people that we were working with. And we've been able to put out several reports which have served the Foundation well. They have kept the White House and kept the Congress from feeling that they have had to manage the National Science Foundation. Our Director is able to say, in effect, We do our evaluation work, we think we do it credibly, we think we do it respectably, so you, there in the Congress, Congressional staff or White House can leave us alone, and they respect that. They don't feel the need to fiddle around or to make very small decisions. So they give the Foundation a lot of room to manage itself and I think part of the value of the evaluation work is in doing that.

Lars Gidefeldt

Peer Review Evaluation

Introduction

The Swedish Natural Science Research Council (NFR) started performing international evaluations in 1977. Over the years 62 evaluation reports have been published, covering research fields in biology, physics, geosciences, chemistry and mathematics (see Table). The Council has been instructed by the Government to evaluate research supported by the Council, and this is the formal basis for the evaluation procedure.

Evaluations are sometimes done in collaboration with other governmental fund-granting bodies. In this respect the Council has co-operated with the Swedish National Board for Technical Development, the Swedish Board for Space Activities, the Swedish Council for Forestry and Agricultural Research, the Swedish National Environment Protection Board, and the National Board of Universities and Colleges.

Some Characteristics of the NFR Evaluations

The evaluations can be characterized as peer review evaluations. Distinguished scientists of highest international standard are chosen to form the evaluation committees. They are appointed by the programme committees for each of the fields of biology, physics/mathematics, geosciences and chemistry. Each programme committee defines the research areas to be evaluated and organizes the evaluations, normally one per year. Due to the fact that scientific experts nowadays are always chosen from the scientific community outside Sweden, the evaluations provide the Council with independent assessments regarding quality and structure of Swedish research, as well as a kind of "international calibration" of the quality of Swedish basic natural science.

Although the evaluations are closely connected to the grants given by the Council, they normally cover most of the Swedish research in a specific area, due to the Council's position as the sole money-granting body in many research fields. In cases where more than one body is active, co-operation is often sought with the aim to cover all fundamental research within the area under review. In general, views of an evaluation committee are thus assessments of a whole area. Taken together the evaluation reports provide a comprehensive view of the whole field of basic natural science in Sweden.

The major part of an evaluation report concerns individual projects. One smaller but important part, however, deals with more general questions, as the Swedish

university system, personnel, availability of heavy equipment, weakly represented research areas, organization of research and similar issues.

The reports not only scrutinize the research in Sweden, they also give advice to the project leaders as to the aim and direction of their research, and to the Council concerning project support, termination of projects, etc. Often the reports also give valuable comments on the general development of a certain field, on need of concentration of resources, on the balance between theory and experiment and on other questions that are important for the deliberations in the programme committees.

One obvious weakness of the evaluation procedure is that the members of the evaluation committees also, and quite naturally, advocate the area being evaluated. The procedure is not well suited for comparison, on an absolute scale, of the quality and research resources in different fields.

Preparations

As already mentioned, each programme committee defines the scientific area to be evaluated. By pursuing its intentions a programme committee can expect its whole field of competence to be gradually covered during a period of eight to ten years. After that period the cycle more or less can repeat itself.

The number of projects to be reviewed in an evaluation is usually limited to about 15 to 25, and the scientific area is defined accordingly. This is important as the number of projects relates to time set aside for site visits and work within the expert group. It is our experience that up to 25 projects can be reviewed by an expert panel during one week. Our experience is also that in most cases five distinguished scientists can set aside one week simultaneously for the site visits. The maximum number of projects in one evaluation was 35 and probably about the upper limit for this kind of evaluation. On that occasion the experts needed ten days to review the projects.

The members of an evaluation committee, usually five persons, are appointed on suggestions made by the researchers. Generally both experimental and theoretical competence are required in the expert panel. Jointly the experts must cover the whole scientific area. One important feature is that the experts should be non--Swedish and independent of the groups to be evaluated.

With great satisfaction we note that the scientists who are asked to join the committee show a great interest in the task. Obviously they consider the evaluation work interesting. Sometimes even a certain proudness of having been chosen a committee member is evident.

A chairperson (rapporteur), usually a member of the programme committee concerned and who is not active in the research field to be assessed, is also

appointed by the programme committee. The secretary of the evaluation committee is the secretary of the programme committee.

The Council has formulated review guidelines for the evaluations. As stated in these directives of procedure, the evaluation committee should comment on the following points:

- the scientific quality of the research results
- the scientific value of the proposed projects
- the value of the methodologies in use and proposed for use
- the capabilities of research leader and of the staff
- the need for the proposed research positions, equipment, etc.
- the question of increased, unchanged or reduced support, or termination of support

In special cases other relevant points can be included in the review guidelines. An example is the evaluation committee for nuclear physics that was asked to take a closer look at the Swedish accelerator situation.

Before an evaluation committee convenes in Sweden for the site visits, the members have received reports from the scientific groups. The reports are normally structured as follows:

- a summary of achieved scientific results during the last six years
- a list of publications covering the last six years
- a plan for the scientific work during the next three to five years
- a summary of the need of resources for posts, materials and travels, expensive equipment
- a summary of scientific activities in research areas outside the field of evaluation
- a summary of the budget and personnel situation during the last fiscal years
- a list of PhD students and PhD examinations during the last six years
- a summary of scientific co-operation with other research groups in Sweden or abroad

Depending on the size of the group, a report should comprise 5 to 30 pages. In addition the group should also provide the evaluation committee with a maximum of ten different publications.

The research groups are given about three months to produce their reports and submit them to the Council office. The office then forwards the reports to the members of the expert panel, 2-3 months in advance of the site visits in Sweden.

The secretary also provides the expert group with information about the Council, the Swedish university system and other issues of interest. The total material sent to the evaluation committee is extensive, and its weight can be as much as ten kg.

One may then wonder if the experts consider all this material to be relevant. The general impression is, however, that most committee members find the material necessary in order to be able to assess the production of the research groups. The committees have not suggested a reduction in the length of the reports nor in the number of submitted publications.

The roles of the chairperson are at least twofold. He or she has a thorough knowledge of the Swedish university system with all its laws, internal rules and traditions. This knowledge is a necessary background to most discussions concerning the organization of research in Sweden. He or she has also a close connection to the programme committee and can therefore lead the work of the evaluation committee, so that the intentions of the programme committee can be fulfilled.

The main role of the secretary is to organize the evaluation. This rather time-consuming task involves close contacts with the chairperson, the evaluators and all the research groups. The secretary is also expected to inform the evaluation committee about the policies of the Council in all matters of relevance in this context.

The Site Visit

Site visits are an important part of the evaluation procedure. Each research group is given the opportunity to present their work to the evaluation committee. The site visits may also include demonstrations of the experimental equipment. It is essential that the whole research group takes part in the activities. In that way the experts can get a view of the scientific standards of the whole group and especially of the quality of the research students.

A visit to a research group takes one to three hours, depending on its size and activities. The research groups have received detailed instructions in advance on how the visits should be organized. It is then stressed that ample time should be allotted to discussions between the research groups and the evaluators. Here one can add that the experts have received detailed information about the scientific results of the group being reviewed before the site visits, and the scientific presentations during the site visit can therefore be very brief.

The evaluation committee must also devote time to internal discussions. These discussions normally take place during late afternoon and in the evening and cover the groups visited during the day. When consensus has been reached on the assessment of the scientific value of obtained results, presented plans and recommendations for future support, etc., one (or two) of the evaluators will undertake the responsibility of writing a first draft of the relevant section of the evaluation report.

Quite often during the discussions within the committee more general problems are in focus. In the final evaluation report these problems are dealt with under separate headings in the general section.

The Swedish universities are situated in six cities. When all of these universities are visited during one evaluation, the schedule involves a great deal of travelling. However, the time then spent on trains, in taxis, in the air and in departure halls need not be entirely wasted - the discussions within the evaluation committee can go on just about anywhere.

The site visits are organized by the secretary of the programme committee in co-operation with one local organizer at each university.

After the completion of the Swedish tour the evaluators have a meeting, for half a day or a whole day, where all remaining questions hopefully are clarified. During this meeting an agreement on the contents of the general section is reached and the distribution of remaining work among the committee members is completed.

In some areas the expert panel has managed to almost complete their report at their last meeting during the evaluation. The final editing is carried out by the secretary, who then sends the edited report to the members of the committee for final approval.

In other cases the committee has to convene for a second meeting (1 to 2 days) in order to discuss the last version of the report. This version, put together by the secretary, is based on individual contributions from the committee members. The outcome of this meeting is the final report.

Summary of the Evaluation Procedure

An evaluation as described above takes about twelve months to complete, from its initiation to the distribution of the evaluation report. In summary the steps involved are as follows:

1. The programme committee defines the area to be evaluated.
2. The programme committee appoints the chairperson (rapporteur).
3. The research groups suggest members of the evaluation committee.
4. The programme committee appoints members of the evaluation committee.
5. The research groups submit reports to the Council.The reports contain information on the objectives of the projects, results obtained, publications, etc.
6. The secretary forwards the reports to the evaluators.
7. The evaluation committee assembles in Sweden and carries out site visits, including presentations and discussions with the research groups being evaluated (in all usually about one week).

8. The evaluation committee discusses each individual project, compares general impressions and (at best) attains consensus of opinion; divides up further work among its members.
9. Each member of the evaluation group sends the secretary a draft of his/her contribution to the documentation. The contributions are edited by the secretary and put together in a preliminary report.
10. The preliminary report is discussed and the final version produced and, if necessary, verified at a meeting with the evaluation committee.

The Report

The report from an evaluation committee is a collective, not an individual product. This is a strong advantage which adds power to the report. In accordance with Swedish law the report is a public document.

The report consists of two parts. In the first part, the committee gives its general views on the state of research in the field concerned and on the development tendencies. Questions of research organization, the university situation and other more wide-reaching matters are discussed, as well as the need for posts, equipment, and so on. In the latter respect, this part is often a summary of the recommendations made in the second part of the report, which deals with the individual projects.

The report is submitted to the Council. This means that the Council takes note of the report, but it makes no commitment to adopt the views and recommendations given in the report. However, it forms part of the material basis for decisions made by the Council regarding applications for grants and provides one of several contributions to the Council's long-term work of establishing priorities.

The suggestions and recommendations delivered in the report concern not only the Council, but also the research groups as well as the universities and other agencies. To remedy weaknesses pointed out by the evaluation, not only the Council has to take action. Many problems can only be resolved by the research groups themselves and within the universities.

The report is distributed to the members of the Council, to members of the programme committee and to the grantees. As a public document it is available to anyone who wishes to examine it.

The number of pages of a normal evaluation report is about 40. The report is printed in 300 - 500 copies and often requested by scientists, university administrators, officials in other grant-giving bodies and by the Ministry of Education.

Not all scientists are content with the assessments in the report. Some communicate their negative views to the Council. Occasionally the evaluation committee is accused of incompetence or of ignorance. Some scientists contend that too little time was allotted to their presentation and to the discussions. Such

communications are distributed to the Council and to the relevant programme committee.

It may happen, of course, that the experts are mistaken or have misinterpreted some information, which has led to a negative evaluation. This is unavoidable with our present scheme, which does not allow the research groups to comment on the text in the report before it is printed. The evaluation reports should therefore be read with a critical mind. It is quite obvious, however, that the expert panel considers themselves competent for each project they have assessed. It is also quite obvious that consensus in most cases is easily obtained. In almost all cases the views of the committee are therefore well founded.

The scientific community has accepted the evaluation procedure. This is shown in "the evaluation of the evaluations" performed by the Council in 1981 and 1988. Two questionnaires have been sent out to all researchers holding NFR grants. The same questionnaire was used on both occasions. Three of the questions and their answers are cited below.

Question: Should evaluations of this kind be made?

Answers : 91 % said Yes in 1981 and 90 % in 1988.

Question: How was the evaluation made in your case?

Answers :	Good %	Satisfactory %	Poor %
1981:	46	38	13
1988:	56	32	9

Question: How was the evaluation made generally in your subject area?

Answers :	Good %	Satisfactory %	Poor %
1981:	38	41	11
1988:	52	39	6

It appears that the practice of performing evaluations and the method by which they are executed were well received by the scientific community already from the start, and that the degree of confidence which the evaluations enjoy may even have increased somewhat over the past decade.

The evaluation reports are also used by the programme committees when assessing applications for research grants. At least one referee (and often two or three) is appointed to scrutinize each application, and the evaluation reports are frequently used in their work.

In most cases the reports contain few elements of surprise. Generally the views of the referees do not differ much from the views of the evaluation committees. This is a verification that the Council's normal and regular assessments are in line with international standards. A report showing that Swedish research is of good quality, or in some cases even of excellent quality, indicates that our internal peer review procedure works well.

One may then argue that the reports might be of limited value to the Council. This is true in the sense that new sensational information is seldom obtained. Actions such as increased support to excellent projects and termination of projects of questionable value are, however, easier to take, and easier to accept for people concerned, if they are supported by an evaluation report. The reports are valuable to the Council also when discussing new priority areas, the need for new equipment, new positions for personnel, etc. The conclusion is that - in many respects - the evaluation reports are indeed important documents to the Council.

In this context one should also keep in mind that the Government has instructed the Council to perform evaluations, although no format has been prescribed. Presently, there is an increasing interest in evaluations within the whole civil sector.

Costs

The total, direct cost paid by the Council for the recent evaluation of systematics amounts to 300 000 SEK, or 50 000 USD. This also includes a small honorarium to the members of the evaluation committee. The total sum of grants from the Council to this field is 5 800 000 SEK (1990/91). Thus the cost of the evaluation is equivalent to 5 per cent of the grants allocated to this field. If one evaluation is performed every eighth year, the cost of the evaluation is about 0.6 per cent of the total project cost. This particular evaluation involved 22 contract holders encompassing 27 projects.

The time devoted by the panel members to the evaluation procedure can be rather extensive. One evaluator (nuclear physics, 35 projects) estimated that he had used 200 hours for the evaluation work. This sum includes preparation (reading all the reports from the research groups), site visits and committee meetings, as well as writing parts of the evaluation report. The time used by the secretary in

connection with the evaluation work is about one month. This time is distributed over one year, but concentrated on time-consuming events like site visits and editing the report.

Conclusions

The evaluations constitute an independent body of information. They identify the strengths and weaknesses of natural science research in Sweden, its capacities, the areas where resources are inadequate, where research stagnates and where it is developing satisfactorily. In this process, the successes and failures of the Council will also be clarified.

Promising young researchers come to light. In this respect, the site visits play a leading role. The evaluation procedure, as developed by the Council, can be used for a limited number of projects. A whole area of research can be covered by an evaluation only in small-sized countries like Sweden. If the procedure is to be used for larger communities only a part of the scientific work can be assessed.

The evaluations provide the Council with expert judgements concerning the work of the research groups and permit direct comparisons between different projects within a subfield. The international evaluations constitute an effective method for obtaining this important information. Such information, in many cases, cannot be procured by examination of the applications for project grants or by other evaluation procedures built into the system.

On the research implementation level, the strong points and the weak points can be identified in the research field concerned, as well as in the university structure and the organisation of research in general.

The structure of research posts and the co-ordination and concentration of resources are important elements in this connection. The evaluations thereby become valuable tools also for the bodies which support sectoral research, for the National Board of Universities and Colleges, and for others. Moreover, the Government and Parliament obtain independent perspectives.

The ultimate aim of the evaluations, however, is to encourage good research. The positive reception which they have received from the great majority of researchers indicates that this aim is being achieved.

Acknowledgements

The author is grateful to Bengt Karlsson, Carl Nordling and Mats Ola Ottosson for valuable comments.

Table - Evaluations

Area	*Year*
Biology	
Systematics of Phanerogams	1977
Endocrinology, Neurobiology and related fields	1979
Physiological Botany and General Microbiology	1977
Radiobiology and Radioecology	1979
Aquatic Ecology	1980
Ecological Microbiology	1980
Chloroplasts and Photosynthesis	1981
Coniferous Forest Project	1981
Taxonomy	1982
Genetics	1983
Terrestrial Vertebrate Ecology	1983
Zoological Cell Biology	1983
Zoophysiology and Functional Anatomy	1984
Prokaryotic Molecular Biology	1985
Plant Hormone Physiology, Cell Techniques in Higher Plants and Morphogenesis	1985
Eukaryotic Molecular Biology	1986
Chemical Ecology	1988
Invertebrate Ecology	1988
Terrestrial Plant Ecology	1990
Systematics	1990

Physics and Mathematics	*Year*
Atomic and Molecular Physics	1978
Physics of Metals	1978
Experimental Nuclear and Particle Physics	1979
Astrophysics	1980
Theoretical Nuclear and Elementary Particle Physics and Mathematical Physics	1980
Mathematics	1982
Geocosmo- and Plasma Physics	1983
Semi-conductor Physics	1985
Atomic and Molecular Physics	1986
Condensed Matter Physics	1986
Nuclear Physics	1987
Elementary Particle Physics	1988

Geosciences	
Geodynamics Project A	1977
Geodynamics Project B	1977
Marine Geology	1977
Hydrology	1979
Solid Earth Physics and Geodesy	1980
Historical Geology and Paleontology	1980
Physical Geography	1982
Physical Oceanography	1983
Geology and Mineralogy	1984
Meteorology	1985
Quaternary Geology	1986
Hydrology	1988
Historical Geology and Paleontology	1989
Solid Earth Physics, Paleomagnetism and Geodesy	1991

Chemistry	*Year*
Chemical Storage of Energy	1979
Nuclear and Radiation Chemistry	1980
Protein Chemistry and Enzymology	1981
Membrane Biochemistry	1981
Organic and Bioorganic Synthesis	1983
Physical Organic Chemistry	1983
Biophysical Chemistry	1983
Electrochemistry	1983
Analytical Chemistry	1984
Inorganic Chemistry with special reference to Solution and High Temperature Chemistry	1985
Physical Chemistry	1986
Structural Chemistry with Diffraction Methods	1987
Theoretical Chemistry	1988
Solid State Chemistry including Materials Chemistry	1988
Biochemistry, especially Molecular Mechanisms	1989
Biochemical Separation and Analysis	1990

John Rekstad

A Comment on Peer Review Evaluation

Gidefeldt has described the procedures applied by the Swedish Natural Science Research Council, NFR, in evaluations of subdisciplines. NFR has achieved considerable experience in this field and has shown how to balance the various interests involved in an evaluation process. When evaluations have to take place, I believe their way of doing it is suitable. The Norwegian Natural Science Research Council (RNF) also applies a similar procedure during its evaluations.

"The ultimate aim of the evaluations is to encourage good research" (Gidefeldt's conclusion). *Scientific quality* is essential, and may be the only criterion when basic science is concerned. High quality science is relevant by nature.

In order to identify high quality science, evaluation in one or another form is necessary. There is no debate about the method. The only competent, and hence acceptable, way of measuring scientific quality is by using experts in the field, so-called *peer review.*

Still I must admit some resistance to broad subdiscipline evaluations, not against evaluations as such, but as they are used by the research councils as a general procedure to achieve information. And I will give a few arguments for this resistance.

Subdiscipline evaluations are resource demanding. I believe Gidefeldt underestimates the costs when considering only the direct costs for the research council. My experience is that these evaluations cost a lot of time and effort for the scientific groups involved. We are not used to measuring time consumption in the scientific community, but there is no doubt that evaluation processes take a lot of attention and power away from "production". It is therefore fair to ask - do the benefits justify these costs?

One argument often used, I do not think Gidefeldt mentioned it, is that the evaluation process is stimulating for the scientists. I will not comment except by saying that active research groups find other and easier ways to get stimulation.

A subdiscipline evaluation ends in a report which is useful both in internal processes in research councils and in communication between a research council and its surroundings - both the political level and research institutions. Still I think evaluations have had limited influence on *decisions.*

The challenge for research councils is to make choices. Considering the small size and the transparency of the scientific community in a small country like Norway, I doubt the information value of evaluation reports for the *programme committees* in research councils. According to Gidefeldt this is also the case in

Sweden: "There are few elements of surprise in the evaluation reports, they more or less confirm the picture the research council already has drawn on the basis of background knowledge and advice from referees on applications for research grants".

Research councils are battle grounds. Although each member is supposed to act independently, he or she has limited insight into branches of science outside their own fields. Somebody else has to judge scientific quality. The more prestige assembled in an evaluation group, the more weight its statements receive. Therefore, evaluations may be used as a weapon in these internal battles. This reveals a deficiency in the research council system which cannot easily be solved. It is certainly not solved by using more and larger evaluations, that correspond only to a change from "conventional weapons to nuclear weapons". What worries me, though, is that this kind of evaluation presents a new opportunity to postpone difficult decisions.

Since the research councils normaliy will be able to predict the result of an evaluation from their own insight, in my opinion there is only one real argument for evaluations of this kind. That argument is also mentioned by Gidefeldt. "An evaluation report makes it easier for a research council to implement decisions".

The neutral judgement of an international expert group strengthens the political platform and authority of a research council. Properly used, I think evaluation reports are of great value in implementing decisions.

Although there might be exceptions, the general procedure for evaluating subdisciplines could be the following. Subdiscipline evaluations should be limited to cases where a research council, according to its own strategy, wants to make changes, e.g expand or reduce an activity, funding of expensive instruments, initiation of new research programmes, etc. This will certainly reduce the number of evaluations, and at the same time make them much more action-oriented.

Evaluations attempt to place national activity within a discipline relative to the international mainstream and research front. This should not be done without considering other conditions and frameworks; that is simply a question of fairness.

Therefore, I am not convinced about the relevance of the Swedish experiences, as reported by Gidefeldt, for the natural sciences in Norway. One should notice the substantial difference in resources and conditions for science in our two countries. This difference is evident from official research statistics and certainly from the experiences gained in collaborative work.

Several evaluations of natural science subdisciplines in Norway have revealed consistent critiques of certain aspects of the Norwegian support system. Suggestions and advice from expert groups have first of all focused on support. The expert groups have found a lot of scientific talent in Norway, but very little support for these talents.

These recommendations and advice have not been followed up, with only a few exceptions. Observed from a basic science level in a university, the situation for natural sciences has not improved during the six-year period since the first evaluation report was published in Norway. So, returning to Gidefeldt's conclusion - the aim of the evaluation is to encourage good research. It is not evident that this aim has been attained in Norway so far.

James McCullough

A Comment on Peer Review Evaluation

I don't have as well prepared a critique as my colleague, but I do have some remarks. First of all, nothing quite like what has been presented by Lars Gidefeldt is going on in the US. The Swedish Natural Science Research Council is doing a much more intensive evaluation of a small number of projects than is possible in the US. If you think of how large the US is, and how many projects and how many universities in a given field, it would be quite difficult to cover everything like that. I did mention that the Academy of Sciences and the Academy of Engineering did some particular analyses about the status of science in the same fields, where the advances were being made and where they thought things might be going - but they did this without assessing particular individuals or groups.

As I mentioned before, I have just come from consulting with the Hungarian Government. They are setting up their research fund. They had one under their Communist regime. It started about five years ago under their Academy of Sciences which is in the Eastern model, governmentally controlled institutes and so forth where they have made grants for research. They have now made this an independent agency. Their parliament has disassociated it from their Academy of Sciences and I was asked to come in and help them establish their procedures and work through a lot of issues and questions.

The money allocation issue is a very difficult one. You said it hasn't changed much here. It is very hard to change in any area. It is especially hard when you are starting in a country and they don't know really where to start as far as making the allocations among the programs. But, with Dr. Gidefeldt's permission I am going to send his report on to the head of that agency because I think it would be very useful for them to have outside reviewers come in and assess their strengths and weaknesses in the various fields. They have some problems of being a small country, high quality science in many areas. But they have immense problems in terms of the universities versus the Academy and the established people versus the people who had not been established and so forth. And so in setting up their system they are very hesitant to critique proposals or to review proposals. Everybody is very tentative about being too critical, so perhaps some teams that could come in from the outside and look at particular areas could be useful to them to give them a better base.

Secondly, I do want to mention something along these lines that is carried out by NSF, where we have site visits by groups of reviewers. They are generally in connection with very big projects and proposals. I mentioned we have about 28,000

proposals a year and most of those are for support of say one professor or two people with some students for a few years' work. But two or three hundred are for very large projects generally involving big facilities, supercomputers, telescopes and so forth.

Also, in the past several years we have tried to establish what are called Science and Technology Centers, or in our engineering area, called Engineering Research Centers. The term 'centers' is a very fluid one; these are multi-field projects with a team of several investigators from different areas of research. Their mission is to collaborate in areas that need integration; we are trying to overcome the disciplinary structure and work across in a multi-disciplinary way. In engineering we have established about twenty-five of these and in non-engineering areas which started later, only a couple of years ago, there are about twenty of these so far in various multi-disciplinary areas.

The original ones in engineering were set up for five years at about five million dollars a year apiece. In making those awards our director and board decided that each center should be reviewed in three years time to see if they should be continued. So we have a process similar to the one that was mentioned here, of getting site visitors who have no connection with a particular project, but we have a problem in that they represent various fields so they must have a spokesman for biology, a spokesman for computer science and so forth on these particular teams. Frequently people from government laboratories and industry are invited too. These site visit teams have the same sorts of organization as the Swedish teams, a chairman and people picked from the outside and so forth. Ours have a problem in that when they go to visit one of these laboratories or centers the scientists there wish to spend the entire time demonstrating how wonderful the science is and putting on slide shows and presentations, so that the review team consequently has to press to have its time to ask its own questions and to write its own report. So sometimes it gets a little bit out of control in that respect.

But the reports of these committees are very powerful. Six years ago we made the first grants for six engineering research centers. When the time came for their third-year review, two of them were discontinued, because these teams made reports to our director and then to our board which said they were not proceeding as well as they could be and were not integrating the science in ways that had been promised in the proposals, so although they have good people, they are doing good work, they are not achieving the kind of integration or cross-disciplinary work. And they were discontinued. There is quite a message in that. So these reviews are now continuing, each year's group is now being reviewed.

I know of one other type of assessment like the Swedish model and this was not just for basic research. The State of Texas a couple of years ago had asked its education and research board to look at all the programs that had been specified in

legislation. Over the years they had accumulated a great many of these, where it said, for example, this university should have a military history department, this university should have such and such in physics, this university should have such and such in mathematics. There were some forty-eight different programs. They could only assess them one by one though because they couldn't match them with each other. But they asked the same sort of questions, had the same sort of reviews, the same kind of statements and reports. And as a result several programs were discontinued by the legislature because they felt that they weren't performing well enough and others were increased in funding. But this was not in one particular scientific discipline, this was across a wide range of incompatible programs.

Anthony F. J. van Raan

Bibliometric Indicators as Research Performance Evaluation Tools*

* This paper is an adapted and extended version of a paper published in the proceedings of the European University Institute Conference on 'Research Management in Europe Today', Florence, 13-15 December 1990.

Introduction

Rationale for Bibliometric Indicators

Scientific research, often in strong interaction with technology, is undoubtedly a major driving force of our modern society. 'Strategic' choices are on the agenda of government and industry. This strongly enhances the need for comprehensive and well-structured information on science. Indeed, we observe an increasing interest in systematic assessment of important aspects of science (such as structure and development of scientific fields, interaction with technology, research performance, international collaboration, etc.). Economic restraints led to sharpening of choices, within fields of science and between fields. Politicians, policy makers, and even scientists call for 'accountability' and 'value for money', simply because funds for science have to be weighted against those for other societal activities, and also within science priorities have to be set.

Traditionally, information on science was primarily furnished by the scientists themselves. This expertise of 'scientific peers' is mainly related to the assessment of the cognitive state-of-the-art of particular research fields. Science policy and R&D management, however, need assessments in a more organizational and structural sense. Examples are the trends of a country's or an organization's share in the worldwide activities in scientific disciplines, the 'impact' of a country in these disciplines as compared to other countries, size and characteristics of international collaboration, the role of developing countries, the role of basic and applied research in new technological developments, the structure of scientific disciplines and their relations with other fields.

Information of the above type cannot generally be provided by panels of peers, since their expertise concerns mainly a qualitative view. Without any doubt these qualitative assessments are extremely important. But nevertheless, peers have increasingly more problems to assess the many aspects of scientific activities, in

particular in the case of application-oriented, interdisciplinary research, and research with specific social and economic aspects. Furthermore, modern science is characterized by many new and rapid developments, the value of which is not always clear, even to specialists. So there is a need for specific data that cannot be provided by peers. Here science indicators come into the picture. *Not* as a replacement of peer expertise, but as a support tool.

We focus here on scientometric indicators. These are quantitative measures, primarily based on data from published material (in particular from the serial literature and, in the case of applied research, from patents), that represent different aspects of the scientific endeavour in a quantitative fashion. As we focus on literature-based scientometric indicators, we prefer to specify them as *bibliometric*.

The development of quantitative methods and techniques for the assessment of research performance and for monitoring scientific developments has been strongly advanced by science indicators research from the 1970s onward. A recent and extensive overview of the field is given in the *Handbook of Quantitative Studies of Science and Technology* (Van Raan, 1988).

Recently, important developments in technology indicators have also taken place. They are, however, beyond the scope of this paper. For the interested reader we refer to a recent review by Van Raan and Tijssen (1990a). In fact, recent research on science and technology indicators may be regarded as the development of *information products*, based on quantitative methods, and tailored in a 'user-oriented' form.

The use and application of numerical methods to describe important aspects of science and, more specifically, the construction of *bibliometric indicators*, is guided by two main principles:

(1) For which aspects of scientific research are indicators desirable?
(2) Can these aspects be expressed properly in a quantitative fashion?

In this paper, we will try to formulate answers to these questions, while pointing at pitfalls and caveats of indicators.

We distinguish three main types of science indicators (a) size and characteristics of scientific *output*; (b) size and characteristics of scientific *impact*; (c) *structural features* of science. The first two types constitute the core of bibliometric research performance analysis, the third type pertains to bibliometric 'mapping of science'.

Indicators of these three main types can be constructed with help of several specific bibliometric methods and techniques. The basic assumption is that bibliometric methods and techniques are only appropriate for those *performers* (groups, institutes, organizations, companies, etc.) and those subject areas (research fields, subfields) where publications (or patents, in the case of R&D activities) are

the principal carriers of knowledge. This bibliometric starting point has the main advantage that it gives a common base for the whole spectrum of scientific activities. Its disadvantage is that the role of 'written knowledge', in particular in scientific journals, is determined by cognitive, cultural and socio-economic constraints which are not the same for all fields of science, countries or research organizations.

Methodological Principles

Methodologically, we may distinguish *one-dimensional* and *two-dimensional* techniques. The one-dimensional techniques are based on *direct counts (occurrences)* of specific bibliographic elements such as publications and patents. We call these techniques 'one-dimensional' as they are in principle represented by *lists* of numbers. The first two of the three above mentioned main types of science indicators, the performance indicators, are mainly constructed with one-dimensional techniques. For example, the size of scientific output is operationalized by the number of publications. The impact of this published knowledge (impact is considered as an important and *measurable* aspect of 'quality') is operationalized by the number of citations received by publications within a certain period of time. One may make a distinction between short-term impact (citations counted in the first three years after publication) and long-term impact.

By 'characteristics' of output (productivity) or impact we mean specific features of publications or citations. For instance, the scientific productivity of a country may be expressed as the total number of publications, and in that case only the 'size' of productivity is measured. One could, however, also distinguish between publications in applied research and basic research, between 'normal' and review papers, or between papers of different research specialities. Again, the 'size' of the scientific productivity with a particular characteristic can be measured by counting the relevant publications. The same applies for impact: one may distinguish characteristics of the cited and citing publications.

The *two-dimensional* indicators are constructed from *co-occurrences* of specific items, such as the number of times keywords or citations are mentioned together in publications in a particular field of science. This reveals linkages between keywords or between citations (*co-word and co-citation analysis*). With a sufficiently large amount of data, all these linkages combine to 'abstract' structures which can be displayed in two-dimensional space *maps of science*. Both the one-dimensional and the two-dimensional indicators can be constructed on micro (research specialties or groups), meso (larger scientific fields, organizations, companies), or macro (national, international) level.

Application Orientation

We conclude this introductory section by translating the above methodological principles into practical terms. To start with the two-dimensional techniques, in this paper we will show that recently important advances have been made in the mapping of science and technology. This 'cartography' allows for the positioning of research groups, institutes, or organizations on the map of science. For research management this is important information: the map is a comprehensive representation of the structural relations of a specific field, with all its subfields and specialties. The positions of the relevant groups or institutes on the map visualize the role these groups or institutes play in that field. Furthermore, these bibliometric maps can be constructed for successive years, thus representing the temporal developments ('dynamics') of the field, together with the (changing) role of the research groups or institutes concerned.

With respect to bibliometric performance indicators - mainly based on one-dimensional techniques - important advances have also been made in recent years. Not only for the natural and life sciences, but also for the technical and applied sciences, the social and behavioral sciences, and in the humanities, it is now possible to build a *bibliometric monitoring system* in order to diagnose important characteristics of research performance (output, impact, international collaboration, etc.) and trends of these performance characteristics over time. Such bibliometric monitors are useful for institutional research management, but also for (inter)national evaluation committees. Our first, and probably somewhat provocative conclusion is that bibliometrics can give us a much more useful tool for research management than commonly is known (or admitted). For example, the development of bibliometric indicators enables us to find answers to questions like: 'What is the scientific activity (in terms of research output) and its impact, of the different European Community (EC) member states in the field of polymer chemistry, sorted by academic versus business-sector research; what are the most recognized groups, and what are their 'specialties'; in what country is international collaboration strongest (and with what other countries); and what can be said about the influence of these research activities on R&D developments in the field of new materials? And, please, can you give us also the trends of all these aspects over the last five years? And finally, can you make a map of polymer chemistry showing its most important research areas and the linkages with neighboring fields?' Finding an answer to questions of this type is now daily practice at our research centre.

Needless to say, a prerequisite for the application of bibliometric indicators is a thorough knowledge of its possibilities and, in particular, its limitations. In the next sections we will explore in more detail the use of bibliometric indicators as a tool for research management. The bibliometric performance indicators (the one-dimensional techniques) are of particular interest for university research policy. The

mapping (two-dimensional) techniques offer comprehensive information that might be essential to research organizations and R&D management of companies.

In this paper we emphasize methodological and technical basic considerations in the use of bibliometric indicators for research evaluation. Given the remarkable resistance in the academic community to bibliometric analysis - partly based on emotional grounds and partly on damages caused by inappropriate bibliometric analyses - it is essential to conduct bibliometric research very thoroughly, in order to keep this type of research on a high professional standard. Therefore, we would like to stress our enthusiasm for careful bibliometric work, and the wealth of possibilities offered by tools based on these quantitative methods. As our group has very broad experience in bibliometric analyses of different kinds, we are able to provide the reader with interesting examples of the practical use of bibliometric methods and techniques for research performance analysis.

Practical Applications

Research Group Performance Indicators

Pioneering work on the development of one-dimensional research performance indicators has been done by Narin (1976) (mainly 'macro level', i.e., the performance of countries) and, in particular for research institutes ('meso level'), by Martin and Irvine (1983). In this paper we focus primarily on the 'micro level': research groups.

As indicated in the introduction, two important concepts play a central role in the development of bibliometric performance indicators: (1) production of scientific knowledge, operationalized by the number (and type) of publications, and (2) impact of this knowledge (this is considered as an important and measurable aspect of 'scientific quality'), operationalized by the number of citations received by publications within a certain period of time. A distinction can be made between short-term impact (citations counted in the first three years after publication) and long-term impact. The operationalization of the above concepts in measurable terms constitutes a set of indicators called *bibliometric monitors*.

In the first of such studies conducted by our group, bibliometric monitors were constructed for about 200 research groups in the basic natural and life sciences. We recently extended our work to the applied sciences and to the humanities and social sciences. Our work in the basic natural and life sciences, concerns the monitoring of Leiden research groups for a period of almost twenty years (1970 - 1987), covering about 12,000 scientific papers and 100,000 citations to these papers. It constitutes a remarkable record of research performance 'histories'. Data were obtained from the *Science Citation Index* (SCI) of the *Institute for Scientific Information* (ISI), Philadelphia. Data handling was for the major part computerized

with specially developed software. A detailed presentation of this *Leiden Science Indicators Project* is given by Moed et al. (1983, 1985) and by Moed and Van Raan (1988). It is a unique project on a larger scale application (a whole university) of bibliometric performance indicators, and may be considered as an exemplar of the possibilities offered by the present state-of-the-art in indicators research.

As discussed earlier, an important presupposition in the bibliometric approach is that results of scientific work are published in the serial literature (primarily journals). In many of the basic natural and life sciences, publication in the serial literature indeed is the major way of disseminating research results. In the humanities and social sciences, books and reports ('grey literature') are also important carriers of research results, and in the technical sciences again books and reports, but also patents, software, designs, artifacts like prototypes, or even maps. However, recent work in our group by Peters et al. (1988), Nederhof et al. (1989), and by Nederhof and Van Raan (1991) shows that for the applied and engineering sciences, for the humanities, as well as for the social and behavioral sciences, international journals do play an important role in the dissemination of knowledge. From these studies we learned that the applicability of bibliometric indicators for a specific field of science depends, in good approximation, upon the extent to which publication databases and, more in particular, citation databases cover the communication channels used by researchers in that field.

Let us now return to the bibliometric monitor as developed in the Leiden Science Indicators Project. Three indicators form the basis of our monitor-system for research group performance. Figure 1 shows a 'real life' example of these indicators (for one of the Leiden physics departments). The following indicators constitute the monitor:

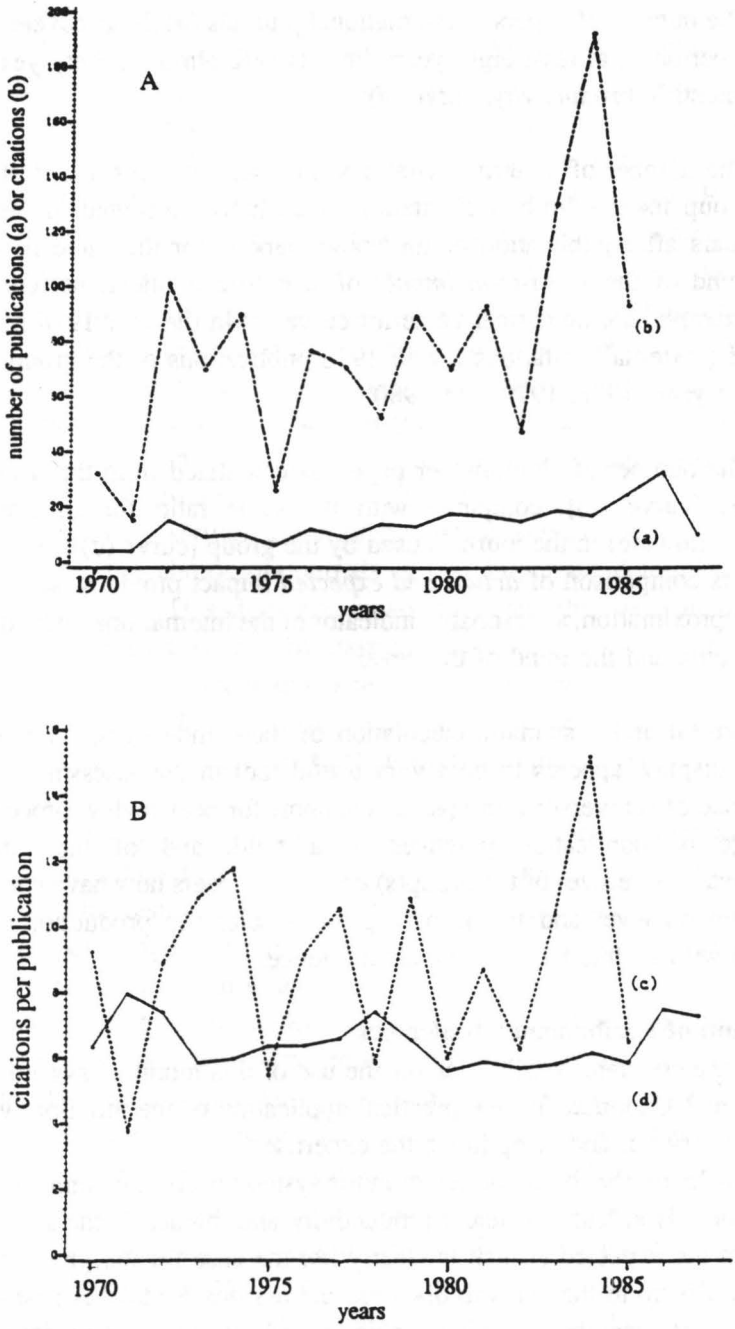


Figure 1: Bibliometric Performance Indicators of a Research Group
part A: curve (a): Number of Publications (Production)
curve (b): Number of Short-Term Citations (Short-Term Impact)
part B: curve (c): Actual (Short-Term) Impact per Publication
curve (d): Expected (Short-Term) Impact per Publication

(1) The number of papers in international journals (as far as covered by ISI) for a period of at least eight years (in this case almost twenty years) trend of scientific *productivity*, curve (a);

(2) The number of 'external' citations (i.e., self-citations and citations by the group itself - 'in-house' citations - excluded) received in the first three years after publication of the above papers, for the same period of time trend of the *short-term impact* of scientific publications, curve (b). For example, the numerical value for curve (b) in the year 1978 is the number of ('external') citations for all 1978 publications of the group received in the years 1978, 1979, and 1980.

(3) The number of citations per paper, as calculated from the data in (1) and (2) [curve (c)], compared with the same ratio for an average paper (worldwide) in the journals used by the group [curve (d)]. We assume that this comparison of *actual and expected* impact provides, at least in a first approximation, a reasonable indicator of the international level of a research group, and the trend of this level.

Such a careful and systematic calculation of these indicators, followed by their graphical display, appears to be a very useful tool in the assessment of research performance of (university) groups, as a support for peer review procedures. With knowledge of publication practices in a field, and of the infrastructural characteristics (like size) of the group(s) concerned, peers now have comprehensive information on level and trends of a group's scientific production and impact, compared with an international impact reference.

Application of Performance Indicators

What are the concrete possibilities for the use of this monitor- system in research management? Undoubtedly, the practical application of the monitor-system is the best way to try out and to optimize the expertise.

With help of the bibliometric monitor-system peers can immediately judge whether there is at least sufficient productivity and impact. If the actual impact is lower than the expected (which is clearly *not* the case for the group presented in Figure 1), it is up to the peers to diagnose the reasons for this. A first explanation is the 'negative' one; the group is not able to perform high-quality research. But we remind that the measured impact is the short-term one. There is a possibility that the impact of the group is of a more longer term, i.e., it takes time before the scientific community recognizes the value of the work. This point could be investigated empirically by determining the longer-term impact. In most cases we

find similar trends for different 'citation windows', and thus we can confine ourselves to the short-term impact. However, there are notable exceptions. Therefore it is important to investigate in the case of low short-term impact *if* differences occur when using longer citation windows. Significant differences in short- and longer-term impact may indicate important aspects of the research conducted in the group, such as the (then) 'ahead-of-time' character of the work (Van Raan, 1989).

If the short-term impact rapidly increases in a recent period (which is the case for the group presented in Figure 1), the peers are confronted with possibly very influential recent work. The trends indicate particular successful publication years, and it is a challenge for peers to explain these findings with excellent Ph.D. work, appointments of new professors and senior researchers, or, indeed, the start of pioneering scientific work. This could be very informative as it is very well possible that peers are not sufficiently aware of specific recent developments or interesting advances made by, for instance, younger scientists.

The often strikingly similar forms of the actual impact and expected impact curves reveal that the choice of journals is one important determinant for the obtained impact value. This does not mean, however, that we can replace the actually obtained impact simply by expected values based on journal impact. Although the forms of the actual and expected impact curves might be similar, the important point is the *difference in absolute values* between actual and expected impact. This difference gives an indication of the impact level compared to an international average.

One can imagine that the availability of bibliometric monitors such as in Figure 1 for all research groups in a university, an organization, or in a country will be regarded as very interesting but 'hot' material. For an example of the comparison of bibliometric performance analysis with peer review (economic research groups), we refer to Nederhof and Van Raan (1991).

We are now extending the Leiden Science Indicators project to all universities in the Netherlands. We have first results for the Agricultural University of Wageningen. This means that we gained experience in the use of bibliometric indicators in application-oriented research fields. Moreover, the social sciences have also been included in the Wageningen study. For some departments intriguing *differences* between bibliometric findings and peer evaluation results have been found. Especially in the cases where bibliometric findings suggest a (much) better performance, it is very important - not least for the departments concerned - to find an explanation for such a difference with peer judgement (Meyer et al., 1991). Meanwhile, a Belgium university commissioned us to conduct a similar bibliometric performance analysis of its natural sciences and medicine faculties.

Another important practical application of bibliometric indicators was our study of six economic research groups in the period 1980-1988 (Nederhof and Van Raan,

1991). These groups participate in a large research programme of the British Economic and Social Research Council (ESRC). Research performance of these groups was compared to the world average by means of the earlier mentioned method of actual versus expected impact. In order to investigate the influence of key scientists (the 'star effect'), we applied a sensitivity analysis to the performance of the research groups by elimination of the papers (and subsequent citations) of such key members. Furthermore, to provide insight into the fields to which a group directs its work, and the fields in which a group has its most important contributions, comparisons were made of publishing and citing journal packets. Similarly, citations to the work of the research groups were analyzed for country and institute of origin. We compared the results of the bibliometric part of this study with those of a simultaneous peer review study (two foreign scientists wrote, as consultants for ESRC, detailed evaluation reports). The bibliometric study yielded clear and meaningful results, notwithstanding the applied nature (economics) of the research groups. Results from peer review and bibliometric studies appear to be complimentary and mutually supportive. In a bibliometrics versus peer review *confrontation meeting*, the participants (i.e., peers, 'bibliometricians', and research council staff) regarded the exercise as most valuable, with lessons for the Research Council both for the future of research programmes and for the form of evaluation used for large awards. We think that outcomes of this 'confrontation' are of general importance with respect to the use of bibliometric indicators. Therefore, the general conclusions of this macro-economic research group evaluation are given, as an example, in the appendix.

A *nationwide* quantitative assessment of research activities allows for (1) a cross-disciplinary monitoring of research group performance for each university, which gives an important tool for the universities to support their own research management; and (2) a disciplinary monitoring on a national (e.g., research council) level. This latter possibility is particularly important in supporting decision-making on future national research activities, stimulation programmes, the establishment of centres-of-excellence, and fruitful international collaboration.

A further possibility is a more refined analysis of the group's scientific impact. Such a detailed specification of 'impact characteristics' may involve the analysis of where the impact (i.e., the received citations) comes from: geographical origin; citing journals; citing authors, groups or institutes and the research (sub)fields or specialties they belong to; the changes over time in these impact characteristics. This type of information is not only useful for research management purposes. Researchers themselves may use these data for tracing patterns of diffusion, use and influence of their research results.

Our practical exercises show that a peer review & bibliometric analysis *combination* is a valuable tool in the performance analysis of research groups. It

also showed that bibliometric analysis never can replace judgements by peers. On the other hand, peer judgement alone will not give sufficient information on important aspects of research productivity and on the impact of research activities. Depending on the quality of both analyses and on the quality of their combination, peer review combined with bibliometric analysis certainly enriches the process of research evaluation in efficiency and effectiveness. We hope that our exercises prove this claim.

Maps of science from practical applications to new epistemological tools?

Science constitutes a complicated, heterogeneous system of activities characterized by many interrelated aspects. Systematic investigation of this network of interrelations, and with that, the structure of science, is an important element of R&D management studies. Nowadays, the enormous and still increasing amount of information on scientific research, as embodied in publications, necessitates a systematic approach to achieve useful data reduction. Large numbers of complex tables are mostly not very useful in this respect. We need new ways of representing the data in order to reveal 'underlying' and until now *hidden* features.

A fruitful approach to solve this problem is the development of 'maps'. The advantages of using such 'cartographical' representations are multiple. A visualization of complex masses of data offers a more complete overview in less time. Furthermore, visual information is more easily remembered. Another very important point is, as indicated above, the *reduction* of information. There is a lot of 'noise' in the enormous amount of data available today. It is a crucial problem to filter the significant features. As we shall see, the mapping techniques developed in our group offer the possibilities to achieve such a data reduction. In other words, a 'cartography of science' not only reformats the data into a specific graphical representation, it also accomplishes data reduction while retaining essential information. The next step is obvious. Maps are not only suitable for depicting a *static* structure. Time-series of maps enables a visualization of *dynamic* features of science, for instance the identification of important changes over time in the development of research fields, or shifts in emphasis of countries, research organizations, or research groups.

Maps of science can be seen as tools for searching, identifying and analyzing structures of scientific activities as reflected by publications. They may point at merging fields of science, emerging new activities, and they offer insight into the position of countries or companies in a field of science. Maps aggregate data in a way no expert, with his or her background and perspective would be able to do. The cartographic approach is, so to say, independent of individual opinions. This is particularly advantageous in the case of broad and heterogeneous research fields.

This does not mean that maps can replace the opinions of experts. A thorough interpretation of science maps requires knowledge about the subject matter of the map, preferably from the 'users'. Therefore, the construction of maps requires a process of interaction between the 'map producers' and the 'customers' to determine the possibilities and the limitations of feasible types of maps.

The advantage of the bibliometric mapping method is the possibility to depict relationships between any combination of bibliometric information elements. Thus, a structure of related keywords (co-word maps), or of related references (co-citation maps), or a structure generated by combinations of keywords and citations can be constructed (Braam et al., 1989, 1991). Each modality refers to another aspect of science and can be applied to different levels of aggregation (varying from R&D groups to entire countries, or entire fields of science).

We briefly summarize the main types of bibliometric maps relevant to our work.

Co-citation maps are based on the number of times two particular articles are cited together in other articles (Small, 1973; Small and Sweeney, 1985; Small et al. 1985). When aggregated to larger sets of publications, co-citation maps indicate clusters of related scientific work (i.e., based on the same publications, as far as reflected by the cited literature). These clusters can often be identified as 'research specialties'. Their character may, however, be of different kind because they are based on citation practices, they may reflect cognitive as well as social networks and relations. Several caveats are involved in this type of bibliometric mapping. To mention a few of the most important: citations only reflect a part of the intellectual structure, and they are subject to a time lag.

A second type of bibliometric mapping is based on *co-word* analysis. Word co-occurrences in a set of publications reflect the network of conceptual relations from the viewpoint of the scientists in the field concerned. These 'co-word' frequencies are used to construct a 'co-word map' which represents research themes in a field of science and their interrelations (Callon et al. 1983, 1986). Co-word analysis is completely independent of citation practices. Main caveats are: words may have other than purely descriptive purposes and their meaning is often context-dependent. The main advantage of co-word analysis is given by the nature of words: words are the foremost carrier of scientific concepts, their use is unavoidable and they cover an unlimited intellectual domain.

In this paper we focus on co-word maps. The main lines of the mapping technique are as follows. For a specific field of science, a representative set of publications is defined. From these publications, all keywords (in the title, or abstract, or the 'controlled terms' given by the database) are collected. Depending on the size of the field and the desired fine structure, the 50 to 100 most frequent keywords are extracted from the entire collection of keywords. For each of these 50-100 most frequent keywords, we determine the number of publications in which

a keyword is mentioned (in the title, or abstract, or in the controlled terms) together with any other keyword. Thus we construct a 50 x 50 (or 100 x 100) word co-occurrence ('co-word') matrix. With the help of multivariate data analysis techniques based on matrix algebra, this co-word matrix can be displayed in two-dimensional space, thus yielding a 'map' in which the structural relations within a research field, based on word relations, are visualized. For further details on the methods and techniques we refer to Van Raan and Tijssen (1990 a, b).

As an example, we present in Figure 2 a co-word map of neural network research and related research fields based on about 20,000 publications in the period 1985 - 1989. Highly related words are located relatively near to each other. Because of the limitations imposed by the two-dimensional representation, one needs an additional 'degree of freedom' to allow for the indication of all related topics (words). Therefore we use connecting lines between related topics. These lines show the skeleton of the structure of neural network research. The clustered words can be regarded as research specialties or *important topics* within the area. We see linkages between different (sub)fields and research specialties, such as topics in biology, cognitive psychology, computer science, and physics. Around the central word (neural network) one observes a biological cluster (upper side left), with related psychological concepts (e.g., connectionism, associative memory) in the (upper) right side of the map. To the (lower) right side there is a large computer science cluster around artificial intelligence and expert systems, developing into pattern recognition and other closely related subjects. To the lower left, one finds important contributions from physics (spin glass). Strong linkages are, for example, visible in the area of visual processing (pattern recognition, picture processing) and in the area of brain research (neurons, brain, synapse). An extensive discussion of our neural network maps, in particular a comparison of a narrative based on review articles with our bibliometric results, is given by Van Raan and Tijssen (1991). A detailed comparison with expert opinions is made by Tijssen (1991).

Maps like this one allow for a 'compact' and surveyable overview of important fields of research. A next step is the identification of the most active research groups in the different parts of the research field, or the positioning of particular research groups on the map. In our opinion, the further development of these bibliometric mapping techniques will supply a very powerful support tool for research management.

But we even try one step further and suggest that bibliometric maps may also have an *epistemological* value, in the sense that they enrich existing knowledge by supplying 'unexpected' relations between specific 'pieces' of knowledge ('synthetic value') or by supplying 'unexpected' problems ('creative value'). The challenging point here is that bibliometric maps may be regarded as *cognitive patterns* resembling stored information in neural nets. In other words, a bibliometric map

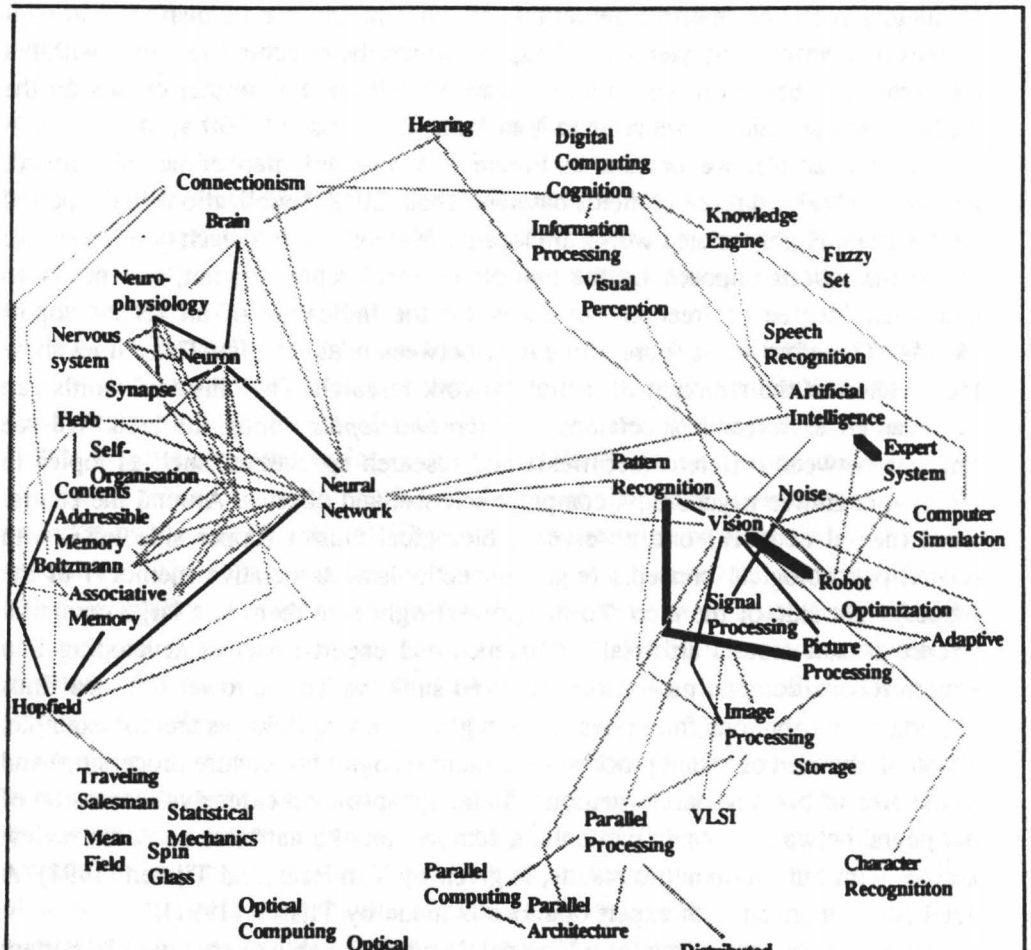


Figure 2. Co-occurrence structure of keywords related to the keyword 'Neural Network', 1985-1988.

Strength of linkages (Jaccard index):

J > .099	━━━
.050 < J ≤ .099	───
.015 < J ≤ .050	·····
J ≤ .015	not shown

represents, in first approximation, the *self-organizing* character of scientific activities in the form of a neural *network-like* structure (and thus our above discussed neural network research map could be regarded as the 'neural network of neural network research'). This paradigmatic metaphor (also developed independently by Ziman, 1991) closely links with our earlier empirical evidence published in *Nature* (Van Raan, 1990) that science can be regarded as a self-organizing system. A further discussion of these rather exotic developments - and not directly related to research evaluation - is beyond the scope of this paper, and therefore we refer to forthcoming work.

Issues of further improvement

Typical Problems in Bibliometric Practice Examples

Returning to the more daily practice of bibliometric indicators, we again emphasize that the most crucial basic assumption in the construction and application of bibliometric indicators is that results of scientific work are published in the serial literature (primarily journals). In the foregoing section we discussed the limitations imposed by this assumption to the applicability of bibliometric indicators. Generally, the applicability of bibliometric indicators for a specific field of science depends upon the extent to which publication databases and, more in particular, citation databases cover the communication channels used by researchers in that field.

Next to problems related to this most central basic assumption, there is a multitude of further methodological and technical problems. For a realistic discussion of potentialities and limitations of bibliometric indicators it is necessary to present a tour d'horizon of these problems and to suggest improvements.

Many of these problems can be solved by further development of bibliometric methods and techniques, but some are more basic. There are several ways to classify problems of science indicators. For instance, problems may be primarily conceptual and methodological, or primarily technical in nature. But in many cases problems are of a mixed type, and after a while a methodological problem may become a purely technical one. We choose a pragmatic approach here. To give the reader an impression of everyday bibliometric practice, we first present a number of typical workfloor problems. It is certainly not an exhaustive list. After that, we define a few clusters of related problems centered around the major question: the applicability of bibliometric indicators.

Science indicators concern different *aggregation levels* of performers (e.g., ranging from individual research groups to a large country), and different aggregation levels of research fields (e.g., from a small research specialty to a major

discipline). A principle requirement is that the subject of analysis (a performer, or a field), expressed as a set of publications, is sufficiently large for statistically significant findings.

An important methodological problem is the *delimitation* of a particular (sub)field of research. In disciplinary databases (like *Chemical Abstracts*) each publication is classified separately with one or more classification codes indicating a specific research field. If no classification code is available, as in the case of the SCI, the definition of specific sets of journals forms another method of classification and delimitation of research fields.

A further methodological problem concerns citations. Apart from the basic problem of the validity of citations as an indicator of 'scientific impact', an important practical problem is how long should one make the time period for collecting citations, the *citation window*. For the calculation of the SCI journal impact (the well-known 'impact factor'), papers published in a two-year period and citations to these papers in the subsequent (third) year are counted. A serious objection against this citation window (at least in the case of research performance measurement) is that the time period between publications and citations is often not long enough for a good measure of impact. The peak in citation scores is field dependent; on the whole, a maximum is reached about 3 years after publication. On the other hand, choosing too long periods makes the results for evaluation purposes less interesting. Here again we have a point of investigation and discussion with evaluators.

A crucial technical problem is the *source* of publication data. There are several possibilities 'hand-made' publication lists (by the analyst and/or by the researchers involved), or professional (commercial) databases such as *Chemical Abstracts* or the *Science Citation Index*. Except for the lists made by the researchers themselves, no source (database) is complete. Important advantages of databases are, however, standardization (to a certain extent), indexing, and selectivity. Selectivity may give rise to problems: most professional databases cover journal publications (however, not necessarily from 'cover to cover'!), often books and conference proceedings (but by far not all books or conferences!), and sometimes reports in 'established' series. For an application-oriented research group with its major output in media not covered by databases (such as occasional reports for government, business sector, international organizations, or publications in special conferences) these international databases may not give a realistic picture of the group's activity. Therefore, the *coverage* of databases is an important point. The *Science Citation Index* (SCI) covers annually about 3500 scientific journals (and a few hundred non-journal publications such as multi-authored books, monographs, etc.).

Some databases include very specific data. For instance, the SCI and her sisters the *Social Science Citation Index* (SSCI) and the *Arts & Humanities Citation Index*

(AHCI), are unique for the inclusion of the reference lists of publications. The 'inverse' of all indexed *references* gives us the citation index. The SCI (as well as the SSCI and AHCI) also includes *multiple addresses* (i.e., more than only the first address of the first author).

A further important point is SCI's annual journal *coverage change*. Annually, about 5 to 10% of the covered journals is changing. These changes are partly due to the journals themselves (for instance, splitting), partly decided by ISI on the basis of the journal's 'income' in terms of citations. This means that for trend analysis, the possible journal coverage change should be investigated carefully.

Another technical problem is the *assignment* (attribution) of publications to a specific group, institution, country, etc. This assignment of publications is generally based on the addresses in the publication as given in the database. Addresses may give serious problems, in terms of completeness, unifications, changes of institutional names, etc. There may even be errors concerning country names. A further problem arises: how do we account for the contribution of co-authors? In the case of 'all-author counting', a paper with more than one author counts as one 'full' paper for all authors. In the case of 'fractional counting', each paper is divided among the contributing authors (or countries). A completely 'justified' fractionalization on the basis of the 'role of the co-authors' is not a very realistic procedure. In our opinion, equal counting of all authors is in most cases the best solution. The consequence that the sum for all authors will be more than the total number of publications can be regarded as, in most cases, a minor technical point.

Main Clusters of Problems and Main Lines for Improvement

Whether a set of bibliometric indicators allows for an answer to questions of R&D managers and policy makers, is critically dependent on the limitations of the indicators, both theoretically as well as practically.

After the above confrontation with a few typical bibliometric workfloor problems, we now define a few clusters of related problems. These clusters partly overlap, and they will not be discussed here exhaustively. Their main characteristics, however, will give the reader a clear overview of the most crucial problematic issues in the use of bibliometric indicators for research evaluation and research management.

The stronger the interaction between users and producers of science indicators, the finer and more precise the aspects of science to be analyzed, the more emphatically basic problems will present themselves. We first mention a cluster of central methodological problems concerning validity and reliability in relation to applicability. The practical problem of *applicability* ('will the constructed indicators meet the needs of the users') approaches closely the question of *validity* ('do the indicators measure what they are designed for; for instance: measuring 'impact' is

not necessarily the same as assessing 'quality'). *Reliability* pertains to the confidence in the numerical values of the indicators: will repeated measurements yield the same results? What is the influence of differences in the way similar or closely related databases are used for the construction of indicators? A well-known recent example is the debate about the question: 'Is British Science declining or not?' (Anderson et al., 1989). This problem arose from different manipulations of one and the same database (SCI).

A second cluster of methodological problems pertains to questions next to the basic problems of validity and reliability. We here mention *accuracy*, and questions related to the statistical significance of calculated numerical values and, in particular, of trends of indicators. Furthermore we have the problem of *relativity*. Bibliometric indicators are not strictly normative. Therefore the question arises: are these indicators comparable to specific standards? There is no theoretical reference to give us an idea about what a high or a low value of a particular indicator means. In practice, one generally compares the indicator value with values found in earlier measurements (preceding periods of time) or with values of other 'performers' (e.g., other research groups, other countries). An example of this practical solution is our comparison of 'expected' versus 'actual' impact.

Durability is a further interesting point. For instance, 'short-term impact' indicators (as discussed in this paper) will not necessarily give an assessment of the 'enduring value' of scientific work.

Finally, we mention a cluster of problems with a more 'technical' nature. First, *collectibility and workability*: Can the data needed be collected? Has the analyst sufficient expertise to carry out thoroughly the many complicated data collection and data manipulation tasks? This latter point relates to the *accountability* of bibliometric indicators. To develop useful indicators, a high level of sophisticated computerized data handling is necessary. Only then, can further indicator work be done on a reasonable economic base. For instance: once specific research databases (mostly 'cleaned' and 'extended' versions of commercial databases) and advanced software packages have been developed, studies such as the Leiden Indicators Project can be done much more efficiently, even on a larger scale (e.g., nationwide).

How can we tackle the above problems? We have to proceed along several lines. First, we must continue the basic methodological work. Indicators research should remain part of quantitative studies of science and technology, and can never be seen as a pure consulting service type of work. Basic research on the underlying assumptions of science indicators is a prerequisite for further advances. The development of science indicators therefore cannot be isolated from studies of science in general. Thorough studies of publication and citation habits in the many different fields of science, in relation to the perception by scientists in these fields with respect to performance (productivity, quality) are needed to supply the

necessary empirical knowledge. Furthermore, it is of major importance to develop analytical procedures for the 'delimitation' of scientific fields in an accurate, systematic way, and to operationalize this delimitation in a bibliometric (if possible) framework.

Mathematical research will be necessary to tackle problems of statistical significance of bibliometric indicators. In particular, the non-Gaussian (skewed) distribution of bibliometric data, such as citations, necessitates the development of new statistical procedures. Mathematical research will also be necessary to improve the mapping of science. In particular, an optimization of multivariate data analytical methods and techniques will be necessary to compare maps of successive periods of time. Further research on the meaning of co-citation and co-word maps, and the relation between such different abstract 'representations' of science has to be done. In particular, the interpretation of the maps by scientists is of crucial importance. This does not mean, however, that scientists should recognize immediately each feature of the map, otherwise the maps would not offer an 'added value'. Our point is, that the 'best possible bibliometric map' must be developed on the basis of further methodological and technical improvements in bibliometric research, in *strong interaction* with the 'customers'.

Conclusions and Recommendations

The training of research managers must include an introduction to new methods of R&D evaluation. Science indicators, and in particular bibliometric indicators, offer exciting possibilities to get comprehensive, to-the-point information on important aspects of scientific development and, in particular, research activities. They are quantitative measures of important aspects of scientific research performance, knowledge transfer and knowledge diffusion, the linkage between science and technology, the structure of scientific fields and the changes in structure over time, international collaboration, etcetera. Needless to say, the evaluation of R&D activities and the use of indicators are closely connected.

Research on science indicators is part of the field of quantitative studies of science and technology. The demand from governmental and international science policy research organizations (universities, research councils), and R&D management is a continuous driving force for the further development of science indicators. Sometimes policy makers and R&D managers want to know everything, and as quickly as possible. Sometimes indicator makers promise too much. This situation may become a danger in the development of valid, useful science indicators. Science is a very complicated system of knowledge production and knowledge exchange. The use of empirical methods in which sophisticated data collection and data handling techniques play a substantial role, is undoubtedly a prerequisite for the advancement of our understanding. Basic research on the underlying assump-

tions of science indicators is another, equally important prerequisite. Research on science indicators, therefore, cannot be isolated from science studies in general.

In our opinion, quantitative indicators based on bibliometric methods can be used successfully in the assessment of research performance, and, more generally, in R&D management and science policy, provided (1) that the presuppositions, on which the indicators are based, are clearly articulated, (2) that these indicators have a sufficiently sophisticated methodological and technical level, (3) that they can offer a variety of 'customer-relevant' information, and (4) that they should enable us to filter significant 'signals' from a large amount of 'noise'.

A continuing interaction between 'makers' and 'users' will undoubtedly enhance the quality of bibliometric indicators. It is, in fact, a prerequisite for further new, exciting, and, above all, useful developments. One of these developments may be the 'epistemological potential' of bibliometric mapping, i.e., its value as a means of advancing knowledge in addition to the knowledge it is based upon. This surplus value may be found in 'synthetic' or 'creative' elements. The first type is related to the discovery of new relations between specific pieces of knowledge, the latter type is related to the discovery of new problems which demand priority in solution. This epistemological potential is strongly related to the idea that science can be conceived as a 'self-organizing system' in the form of a 'neural network-like' structure of which the bibliometric map is a first-order-approximation.

Appendix

A Practical Example: General Conclusions of the Peers on the Bibliometric Analysis of Six Economic Research Groups (from Nederhof & Van Raan, 1991)

The peers felt that the bibliometric study provided much more than a simple measurement of the quantity of work done by the research teams: "It helps a lot in evaluating the quality of the work done". One must, however, be very careful in interpreting the data. One peer wanted to stress a few important points, mainly in relation to the role of journals. We indicate these points, followed by our comments.

First, the *quality of the journals* in which the papers have been published is an important criterion, at least as important as the number of published articles. Second, the number of citations need not be a good indicator of impact, because the *researchers* quoting the paper may be of different quality. Furthermore, the size of the audience will differ from one topic to the other. Technical papers are likely to be less often read and quoted than less technical ones, especially if they are of average quality. Concerning this second point, we remark that citation analysis is confined to only those citations given by scientists publishing in SCI-covered journals. Thus, these citations will be given, on the average, by, at least reasonably qualified researchers. The 'size of the audience' (mainly field-dependent) is taken

into account (but of course never 'completely') by the 'expected' citation level indicator. As technical papers will be often published in the more technical journals, (dis)advantages (with respect to citation scores) of this type of papers are at least partly taken into account by using the expected value.

Third, comparison of actual with expected impact does not entirely solve the *level problem*: it may be preferable to have a number of actual citations below the expected level in a good journal, rather than the opposite in a low quality journal. Concerning this third point, we note that publishing in high quality journals often leads to a higher impact than publication in a journal of lesser quality. Recent results (our Wageningen study) indicate that at the *research group level* the actual impact is determined by the journal for about 50% (a more detailed analysis will be presented by Meyer et al., 1991). We have also compared both the short-term impact levels and the 'expected levels' of the six groups, which makes it unlikely that groups 'suffer' because of publication in good quality journals.

Fourth, a further specification of impact in terms of *foreign impact*, e.g., the ratio of US (or other foreign) to UK citations as performed in this bibliometric study, was regarded by the peers as an interesting indicator of quality.

Returning to the more general role of bibliometric analysis, the peers stated that the number of citations is strongly related to the *choice of a research topic*. For example, the peers felt that one of the papers in this evaluation was highly cited just because a lot of work has later been done in the same area. In our opinion, however, this is a too negative attitude as it is certainly an important aspect of scientific quality to be in the lead of a new development.

According to the peers, the *timeliness* issue posed an important problem. Some of the articles of which the impact was measured by citation analysis, were published before the current grant periods started, and must have been written well before them. For some groups this means that a part of the total work was done under previous ESRC grants, or independently of ESRC grants. Therefore, it might be useful to isolate the productivity of the grants themselves, rather than that of individual researchers, by concentrating on publications emanating directly or indirectly from ESRC financed research. By request of ESRC, we considered a rather long period (1980-1988). In this way, the bibliometric analysis covered the time before and after the award of the grant.

Last but not least, the peers could not avoid discussing the negative impression given by *downward bibliometric trends*. According to one, this finding probably reflects the declining audience, which may be the result of the success of the initial research programme. The peers emphasized that downward trends do not necessarily imply a declining quality of the work. The joint research effort by the ESRC is unique in that it brings together people from different research centres and forces them, to some extent, to cooperate. Because many of the initial objectives have

been achieved, economists from outside the research programme may look for new ideas. Basically, this argument seems to imply that the research was primarily (and successfully) *applied* in nature, and did not generate primarily new ideas, and therefore, citations declined. In our opinion, the bibliometric analysis was successful in revealing this development. The peers found it difficult to interpret the declining trends in citations for several of the groups. In the more dramatic cases, the negative trend mainly originates 'artificially' after an earlier peak by the 'star effect'. The reputations of these principal researchers were built before the preceding grant periods commenced. There were no such single high-impact papers produced under the 1983-1987 grants, hence the declining trends. Since the start of the new grant periods, the publication trends of the groups have not tended downward.

One peer noted that in the bibliometric analysis the 'forward-looking' expected 3rd-year impact of the 1987 and 1988 publications was not computed. He argued that for future assessments of this type, 'total' and 'per article' indexes of expected future citations might provide a useful way of summarizing the current journal publication performance of the groups. The totals are a measure of the strength and breadth of the 'dissemination' efforts, at least for the range of publications covered by the study, and the per article expected citations would indicate the extent to which the papers are being successfully placed in heavily-cited journals.

The peers stressed a problem of measurement bias in the comparison of different groups. It is particularly problematic in the *evaluation of applied research* - especially that focussing on macro-economic properties and performance in a single country - by publication and citation counts from international journals primarily devoted to more theoretical, and hence more tradeable topics. In particular, it might be expected that the work of one specific group would be more citation prone, given its concentration on easily transferable methods and techniques. The bibliometric analysis, however, revealed that this group was not very often cited. For applied work, especially research aiming at *policy-relevant, quantitative results*, journal publication and impact measurement by citation analysis will provide only a tangential measure of research quality and impact. Given their relative interests, groups may differ in the relative weights they attach to the applied and theoretical components of their research, and hence to the choice of publication channels, and, with that, the relevance of bibliometric analysis.

According to the peers, bibliometric analysis does provide a useful check and sidelight on conventional content-based peer review. Excessive reliance on these measures, however, needs to be avoided. Research groups ought not be encouraged to think that their rankings will be closely dependent on citation rankings, since the resulting re-direction of research and publication efforts might well be at the

expense of the goals for which a research programme was established in the first place.

The idea of having the peer review and bibliometric analysis proceed independently, at least to the stage of a report reviewable by a committee, is probably a good one, so as to increase the range of independent information available to committee members. It might, however, be useful to supplement these primary reports with revisions or commentary based on discussions among the consultants.

This confrontation of peer review and bibliometric analysis can be summarized (but not exhaustively!) in the following inventory:

- *Journals* are an important but certainly not a 'complete' predictor of future impact;
- The *small numbers* of publications in indexed journals mean that only large differences are significant;
- Even if there is a 'genuine' downward trend in citations, this does not necessarily indicate a downward trend in *usefulness* to policy makers or in other types of output such as professional training of younger and visiting economists who worked in the research teams. In other words the bibliometric approach cannot pick up the 'enlightenment' of policy makers by the produced scientific work, nor its importance for training; it gives a measure of academic impact;
- It appeared that especially theoretical work receives most impact, and applications come off worse in this respect. Therefore, application-oriented work probably was more influential than the impact analysis would suggest. Bibliometric approaches miss the 'comfort' given to policy makers by *applied research;*
- For at least three groups a major source of impact, as measured by citation analysis, was a publication *not related* to the research programme on macro-economic modelling;
- The bibliometric approach was considered to be important in 'agenda setting', but the results of the study should certainly *not have a dominant weight* in policy recommendations. Given the cost of this bibliometric study (less than k£ 10) in relation to the size of the awards, the exercise is certainly worth repeating in the future but perhaps in a modified form;
- The timing of the *citation window* (e.g., short-term versus long-term) was considered to be crucial;
- A *geographical breakdown* of the received impact, which is shown to be possible in this study, is of interest as a special indicator of scientific quality;

- A practical conclusion of this exercise is that there was no clear advantage in having qualitative or quantitative review work prepared first. Only *slight changes* to the present exercise would be necessary for an 'ideal' balance;
- Bibliometric analysis would be particularly helpful at the outset of awards, it gives a valuable guide to the *track record* of applicant groups;
- The value of bibliometric studies is in helping to formulate pertinent questions, but 'literal' and 'mechanical' application is not appropriate. Bibliometric analysis can *never replace peer review*. Integration with peer review remains essential. Under these conditions, bibliometric analysis is a useful support tool.

References

Anderson, J., P.M.D. Collins, J. Irvine, P.A. Isard, B.R. Martin, F. Narin, and K. Stevens (1988). On-line approaches to measuring national scientific output: a cautionary tale. *Science and Public Policy*, 15, 153-161.

Braam, R.R., Moed, H.F., and Van Raan, A.F.J. (1989). Comparison and Combination of Co-citation and Co-Word Clustering. In: *Science and Technology Indicators. Their Use in Science Policy and Their Role in Science Studies*, Edited by A.F.J. van Raan, A.J. Nederhof and H.F. Moed, pp. 307-337. Leiden: DSWO Press.

Braam, R.R., Moed H.F., and Van Raan, A.F.J. (1991). Mapping of science by combining co-citation and word analysis, I: Structural Aspects; II: Dynamical Aspects. *Journal of the American Society for Information Science* (in press).

Callon, M., Courtial, J.-P. Turner, W.A. and Bauin S. (1983). From Translations to Problematic Networks: An Introduction to Co-word Analysis. *Social Science Information*, 22, 191-235.

Callon, M., Law, J. and Rip, A. (Eds.) (1986). *Mapping the Dynamics of Science and Technology*. London: MacMillan Press Ltd.

Martin, B.R. and J. Irvine (1983). Assessing Basic Research: Some Partial Indicators of Scientific Progress in Radio Astronomy. *Research Policy* 12, 61-90.

Meyer, R.F., A.J. Nederhof, H.F. Moed, and A.F.J. van Raan (1991). Performance and utility indicators for agricultural and veterinary research in the Netherlands. Report to the Netherlands Council for Agricultural Research (draft version).

Moed, H.F., Burger, W.J.M, Frankfort, J.G. and Van Raan, A.F.J. (1983). *On the Measurement of Research Performance: the Use of Bibliometric Indicators.* Centre for Science and Technology Studies, University of Leiden, Leiden.

Moed, H.F., Burger, W.J.M., Frankfort, J.G., and Van Raan, A.F.J. (1985). The Use of Bibliometric Data for the Measurement of University Research Performance. *Research Policy*, 14, 131-149

Moed, H.F. and Van Raan, A.F.J. (1988). Indicators of Research Performance Applications in University Research Policy. In: A.F.J. Van Raan (ed.), op cit., pp. 177-192.

Narin, F. (1976). *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity.* Washington D.C.: National Science Foundation.

Nederhof, A.J., Zwaan, R.A., De Bruin, R.E. and Dekker, P.J. (1989). Assessing the Usefulness of Bibliometric Indicators in the Humanities and the Social Sciences: A Comparative Study. *Scientometrics*, 15, 423-437

Nederhof, A.J. and A.F.J. van Raan (1991). Bibliometric Analysis of Macro-Economic Research Groups: A Comparison with Peer Review (draft version).

Peters, H.P.F., Hartmann, D. and Van Raan, A.F.J. (1988). Monitoring Advances in Chemical Engineering. In: *Informetrics* 87/88, edited by L. Egghe and R. Rousseau, pp. 175-195. Amsterdam: Elsevier Science Publishers.

Small, H. (1973). Co-citation in the Scientific Literature: A New Measure of the Relationship Between Publications. *Journal of the American Society for Information Science*, 24, 265-269.

Small, H. and Sweeney, E. (1985). Clustering the Science Citation Index Using Co-Citations, I: A Comparison of Methods. *Scientometrics*, 7, 393-404.

Small, H., Sweeney, E. and Greenlee, E. (1985). Clustering the Science Citation Index Using Co-Citations, II: Mapping Science. *Scientometrics*, 8, 321-340.

Tijssen, R.J.W. (1991). Internal and external representations of science: comparing bibliometric maps with expert data. Centre for Science and Technology Studies, Report CWTS 91-04, Leiden.

Van Raan, A.F.J. (ed.) (1988). *Handbook of Quantitative Studies of Science and Technology*. Amsterdam: Elsevier Science Publishers.

Van Raan, A.F.J. (1989). Evaluation of Research Groups. In: *The Evaluation of Scientific Research*. D. Evered and S. Harnett (eds.), p. 169-187. Chicester: John Wiley.

Van Raan, A.F.J. (1990). Fractal Dimension of Co-citations. *Nature* 347, 626.

Van Raan, A.F.J. and R.J.W. Tijssen (1990a). An Overview of Quantitative Science and Technology Indicators Based on Bibliometric Methods. Technology/ Economy Programme (TEP), OECD, Report Nr. 27769, Paris.

Van Raan, A.F.J. and R.J.W. Tijssen (1990b). Numerical Methods for Information on Aspects of Science: Scientometric Analysis and Mapping. In: *Perspectives in Information Management*, Vol.2. Ch. Oppenheim, J.-M. Griffiths, and Ch. L. Citroen (Eds.). London: Butterworths. To be published.

Van Raan, A.F.J. and R.J.W. Tijssen (1991). Bibliometric Mapping of Neural Network Research. To be published.

Ziman, J. (1991). A neural net model of innovation. *Science & Public Policy*, 18, 65-75.

Per O. Seglen

A Comment on Bibliometric Indicators as Research Performance Evaluation Tools

In his review, van Raan refers to the "remarkable resistance in the academic community to bibliometric analysis", and ascribes it to "emotional grounds" and to "damages caused by inappropriate bibliometric analyses". I feel that scientists should be given more credit than that; bibliographic references are, after all, a major scientific tool, the use and abuse of which is familiar to all of us. We thus know from our own daily practice that references are used for the purpose of documentation rather than to reward our colleagues or to produce an accurate historical record. Journal space limitations allow us to cite only a small fraction of the relevant literature; a selection therefore has to be made that provides ample room for both randomness and bias. The frequent experience of not being cited when we ought to be further contributes to our quite unremarkable, healthy scepticism about the soundness of using citations as a basis for scientific evaluation.

The problems with the use of citation data for evaluation can be broadly classified into four major groups:

Conceptual problems

Citational impact (citation frequency) does not equal quality, and may not even correlate particularly well with quality. Citations refer to the *use* of scientific work, hence they are primarily a measure of utility. Citation of *methods* may provide a case in point: a method of general applicability can become very widely cited (provided it is reasonably useful), whereas a method developed to solve a particular scientific problem may not be cited outside the group working on that problem, no matter how ingenious the method is. Method citations also carry some peculiarities: while analytical *procedures* are usually referred to by citation, analytical *tools* are not. Thus Lowry's very general protein determination method, which is not even very good, is cited something like 10,000 times each year, while the discoverers of widely used inhibitors of protein synthesis like puromycin and cycloheximide are never acknowledged: this is the syndrome known as "obliteration by incorporation". Citedness can thus clearly be dominated by factors other than scientific quality (1); some of these factors are listed in Table 1.

Field effects

Citation frequency is determined by the contents, relations and dynamics of the scientific field. Different scientific fields can have widely different average citation rates which can be regarded as a technical property of the field rather than as a reflection of the scientific quality of that field. For example, van Raan has previously shown that biochemists on a short-term basis are cited four times as often as mathematicians, simply because biochemists use more references per article, and tend to refer to more recent work (2). Papers in basic medical science may be cited three to five times as often as papers in clinical medicine, because there is largely a one-way citational relationship between basal and applied sciences (3,4). There are also other field factors, summarized in Table 2. It is quite possible that these field-specific citational characteristics may extend even to microfields (scientific specialties). A citation analysis of subsections and defined subfields within two major scientific journals thus suggested a large degree of heterogeneity (Table 3), i.e. the activity profile of each scientific group may define a unique *citation aura*, which determines the average impact independently of quality. It is obvious that adequate correction for such individual citation auras cannot be made, thus making it impossible, in most cases, to distinguish between field effects and quality effects at the level of individual scientists or research groups.

Table 1. **Problems of Reference Selection**

1.	Utility, not quality as primary criterion
2.	The citation probability is low
3.	Incomplete citational coverage
4.	Obliteration by incorporation
5.	Citation of secondary sources
6.	Argumentative citation
7.	Flattery (of potential editors/referees)
8.	Convention (e.g. in methods citation)
9.	Reference copying
10.	Self-citation and "in-house"-citation

Table 2. **Field effects**

1.	Reference immediacy
2.	References per article
3.	Field dynamics (expansion/contraction)
4.	Interfield relations (e.g. basal/applied)
5.	Microheterogeneity (citation aura)

Table 3. **Citation frequencies in different biochemical subfields.** Citations in 1987 to all articles in a sample of *Biochem. J.* and *J. Biol. Chem.* issues from 1984.

Biochemical Journal	
All articles (vols. 211-214)	4.17 ± 0.45 (446)
Cellular Aspects (vols. 212 & 214)	5.00 ± 0.76 (252)[a]
Molecular Aspects (vols. 211 & 213)	3.10 ± 0.27 (194)[a]
Subsections	
Cell surfaces and receptors	7.35 ± 3.26 (34)
Metabolism, regulation and control	5.08 ± 1.09 (142)
Peptide and protein structure	4.25 ± 0.90 (44)
Membranes, transport, bioenergetics	3.59 ± 0.54 (34)
Carbohydrates	3.18 ± 0.65 (22)
Enzymes and enzyme kinetics	2.44 ± 0.30 (81)
J. Biol. Chem.	
All articles (issues 11-14)	6.93 ± 0.41 (403)
Subfields (title keywords)	
Calcium	14.06 ± 3.67 (16)[c]
Receptors	13.57 ± 3.17 (28)[d]
Protein kinases and phosphatases	6.92 ± 1.53 (12)
Mitochondria	5.21 ± 0.84 (14)
Plasma proteins	5.00 ± 1.15 (40)
Other enzymes	4.74 ± 0.52 (85)
Bacteria	4.30 ± 0.65 (37)[b]
Plants	1.71 ± 1.04 (7)

[a]Cellular and Molecular Aspects significantly different at the 95% confidence level. [b,c,d]Significantly different from journal mean (all articles) at the 98%, 99.5% and 99.9% confidence level, respectively.

Choice of evaluation parameter

Van Raan suggested that the journals in which articles are published are representative of the scientific field, and that correction of field effects may be achieved simply by dividing the citation frequency of the article by the mean citation frequency of the journal (the *journal impact factor*). Unfortunately the situation is not as simple as that. Within my own field - biochemical cell biology - there are hundreds of journals which are equally representative of the field, but which vary in impact from zero to twenty (as compared to a mean value for the field of about three). If I knew that my grant applications were to be evaluated on the basis of van Raan's relative impact factor, I would of course publish my papers in low-impact journals to receive a high score. However, I might be fooled: my grantors might have changed their mind, and instead listened to those who think that the quality of scientific work can be measured by the quality of the journal in which it is published. By thus facing the journal impact factor as an evaluation parameter, I would have been better off publishing in high-impact journals. It may be (and has been) argued that high-impact journals are preferable in any case, because an article in a high-impact journal is automatically more cited than an article in a low-impact journal, but correlation studies at the single research group level have failed to provide support for this contention (4,5). The journal impact is, furthermore, not very representative of its component articles: the individual articles differ enormously in citedness. The most cited half of the articles account for almost 90% of the citations (6,7), and the majority of the articles are more than 50% away from the journal mean.

Table 4 illustrates how the choice of evaluation parameter can determine the outcome of an evaluation. Nine research groups have been ranked on the basis of real citation frequency, journal impact or relative impact, and as can be seen the three ranking orders obtained are very different. If any of these bibliometric parameters are to be used for evaluation, scientists should at least be given due warning some years in advance, to get time to choose a score-optimizing publication strategy. Whether science is served by directing the effort of scientists towards impact optimization rather than towards scientific quality can of course be questioned.

Table 4. **Effect of bibliometric parameter choice on evaluation result**

Citation Frequency (cit./year/article) (CF)	Journal Impact (JI)	Relative Impact (CF/JI)	Rank order on the basis of Citation Frequency	Journal Impact	Relative Impact
10.95 ± 3.39 (21)	3.06 ± 0.45	3.58	1	3	1
6.93 ± 1.25 (29)	4.40 ± 0.58	1.58	2	1	5
3.87 ± 0.89 (15)	1.32 ± 0.16	2.93	3	8	2
3.80 ± 1.16 (15)	2.65 ± 0.69	1.43	4	5	6
3.79 ± 0.84 (24)	2.32 ± 0.60	1.63	5	6	4
3.38 ± 0.87 (13)	2.93 ± 0.34	1.15	6	4	8
2.13 ± 0.74 (8)	1.60 ± 0.33	1.33	7	7	7
2.05 ± 0.34 (22)	1.01 ± 0.15	2.03	8	9	3
1.89 ± 0.84 (9)	3.83 ± 1.13	0.49	9	2	9

From nine biomedical research projects, all journal articles 1976-82 with the project leader as first author were analyzed two years after publication with regard to citation frequency (CF) and the corresponding Journal Impact Factor (JI) of the journal in which the article was published. The relative impact (CF/JI) has also been calculated, and a ranking of the groups on the basis of each bibliometric parameter is presented. CF and JI values are given as the mean ± S.E. of the no. of articles indicated in parentheses. Modified from (4).

Accuracy problems

Even if we were to accept citational impact as a valid evaluation parameter, it is not obvious that it is technically suitable for the purpose. Are, for example, citation frequencies sufficiently stable to be representative? If single groups are examined, it becomes obvious that article citation frequencies distribute extremely heterogeneously, necessitating a very large material to establish statistically significant differences between groups (7). In the example given by van Raan (his Fig. 1), the analyzed group would seem to lie well above the expectation level, but the variability is so great that it does in fact take ten years before the difference becomes statistically significant. On the individual group level, citation data would therefore seem to be unsuitable for most practical purposes on a purely statistical basis.

My conclusion is thus that bibliometric methods cannot be used for evaluation of individual research groups. At the level of scientific institutions and departments, field heterogeneity will become an increasingly serious problem, making bibliometric evaluation unsuitable even here. It is probably only at the national level that a given field is large enough to assume statistical homogeneity, making it possible to perform valid bibliometric comparisons between nations within well-defined research fields. However, although bibliometric methods have limited applicability in evaluation, it deserves to be pointed out that bibliometry is a fascinating and rewarding research field in its own right, and that the possibility of asking and answering metascientific questions in a quantitative manner may provide valuable insights into the basic sociology of science.

References

1. MacRoberts, M.H. and MacRoberts, B.R. (1989) Problems of citation analysis: a critical review. *J. Am. Soc. Information Sci.*, 40, 342-349.

2. Moed, H.F., Burger, W.J.M., Frankfort, J.G. and Van Raan, A.F.J. (1985) The application of bibliometric indicators : important field- and time-dependent factors to be considered. *Scientometrics*, 8, 177-203.

3. Folly, G., Hajtman, B., Nagy, J.I. and Ruff, I. (1981) Some methodological problems in ranking scientists by citation analysis. *Scientometrics*, 3, 135-147.

4. Seglen, P.O. (1989) Bruk av siteringsanalyse og andre bibliometriske metoder i evaluering av forskningsaktivitet. *Tidsskr. Nor. Laegeforen.*, 31, 3229-3234.

5. Seglen, P.O. (1989) From bad to worse: evaluation by journal impact. *Trends Biochem. Sci.*, 14, 326-327.

6. Seglen, P.O. (1989) Kan siteringsanalyse og andre bibliometriske metoder brukes til evaluering av forskningskvalitet? *NOP-Nytt (Helsingfors)*, 15, 2-20.

7. Seglen, P.O. (1991) Evaluation of scientists by journal impact. In Weingart, P. and Sehringer, R. (eds.) *Science and Technology Indicators*, DSWO Press, Leiden, in press.

James McCullough

A Comment on Bibliometric Indicators as Research Performance Evaluation Tools

Dr. van Raan in his paper mentions the value of bibliometrics for assessing the performance of research organizations and individuals, notwithstanding all the pitfalls and problems we just heard about. Let me just describe a report that someone on my staff is working on that we expect to publish within a couple of months. At the moment we are trying to sort out methodological problems and also to put it in much more readable form for our non-technical audience. We have attempted to use bibliographic methods as a measure of our own performance as a granting organization, specifically by looking at the question how well has the peer review system made its decisions to sort out those who should continue to get NSF grants and those who should not. So we were looking at the question of do our divisions renew their better grants? And in looking at that do NSF grantees publish in better journals? Do they have highly cited papers in the same journals and so forth?

We picked one division from each of our five research support directorates for political reasons. We have astronomy in the physical sciences and in our computer science directorate we picked out computer research, in engineering, electrical and communications systems, and in biology, molecular biology. And in our geoscience area and I think for reasons I'll mention later this was a mistake, our polar program.

We looked at seventy proposals that had been declined or awarded from each of those divisions and we took the bibliography which is submitted by the researcher when the proposal comes around for review. In that bibliography they are supposed to refer specifically to publications that had been produced on a previous NSF grant. From the seventy proposals in each division (about three hundred and fifty grants altogether) we got altogether fifteen hundred and four papers that acknowledged NSF support.

Computer Horizons in Philadelphia counted the citations for each of those fifteen hundred and four papers and when they found them in a particular journal, they then took the next similar article in the journal that was not supported by NSF as a comparison point. So they created a sort of self-comparing database in this way. They had a reference list of articles supported by NSF, and articles in the same journals not supported by NSF.

The results were that for three of our divisions, namely astronomy, electrical engineering and molecular biology, we saw that in retrospect those divisions had indeed been sorting out the more productive from the less productive researchers

when the time came around to renew their grants. For computer science it was very mixed because one-third of the computer science grantees had not published at all. And we learned later by talking with several of the computer scientists, back to Van Raan's point about not coming to conclusions without talking to the peers, that in their own view their publication practices in that field are extremely sloppy. For instance, they give talks at workshops, and in conversations or papers people refer to those talks for years and they never bother to publish. Also it is a developing field and there is a lot of turbulence in the field. For our polar program that was a very mixed case, and I think the reason is that it is not a coherent field of science. It is an aggregation of various areas where they are working in the Arctic and the Antarctic; earth sciences, biology, social sciences, and so forth, a lot of unrelated fields so we couldn't tell too much from that.

We did find in comparing the average citation ratios from NSF supported papers with other papers in the same journals that for our astronomy programs and computer science and electrical engineering, those papers are cited twice as often as papers supported by agencies other than NSF. So we were supporting the most prolific researchers in these fields. We have a direct comparison there since so much molecular biology is supported by the National Institutes of Health. So that gives us a little bit of an argument again to go to the political system and say that: "Yes, you put a lot into the Institutes of Health and these other areas, but when you put it into NSF you get a lot more return for your money."

Now we also looked at the question of non-citedness. You may be familiar with this little controversy that started in *Science* magazine a few months ago by David Pendelburg, who was saying that quite a proportion of scientific papers are not cited at all, and that therefore a lot of what is done in science is worthless. That shocked an awful lot of knowledgeable scientists. Well, it turned out that he was referring to not only journal articles, but all sorts of letters and notes, and summaries and all sorts of things which you wouldn't expect to be cited. Anyway, after taking those out we did our own comparison and found that in most cases NSF articles were non-cited to a lesser degree.

This is a complicated report, and as I said we are trying to simplify it for our audience which is not only research program managers in our own Foundation but Congressional staff, White House staff, media and so forth. And we did one earlier on behavioral and neural sciences at NSF which had similar outcomes with regard to comparisons with support from other agencies.

We are doing a couple of things now with bibliometrics, but we are not too far along on them. One is a project which was started several months ago. We are looking at the big amounts of money that were put into computer science in the 1980s in those departments to see what effect that had and we are trying to use bibliometric data to tell the impact of that funding over the years. Another that we

are now starting on is also going to have a bibliometric component, that is, in our neuroscience area where we're comparing our role with the National Institutes of Health.

We are not doing too much bibliometric work, but we are starting to incorporate it more and more into our various projects and particularly as we develop sources, contractors and so forth who know how to do this kind of work and to develop our own staff expertise in how to use this kind of work and what the problems are. So I think we are pushing into this area rather crudely and we need to develop our own sophistication.

Ken Guy

Academics and Consultants in the Evaluation of R&D Programmes

Introduction

This presentation sets out to examine the roles of consultants and academics in the evaluation of government-sponsored R&D programmes. In order to understand what we mean by the terms 'academic' and 'consultant', some common conceptions are first described. A distinction is made between different evaluation tasks and the institutions or organisational actors normally expected to conduct them, prior to pointing out a number of factors which are currently making it harder to arrive at any simple correspondence. Critical variables in the choice of academics and/or consultants in programme evaluations are then described in more detail, with actual roles discussed via reference to five specimen evaluations. Finally, lessons from each of these evaluations for the involvement of academics and/or consultants are drawn, together with some very general lessons for the commissioners of evaluations.

Academics versus Consultants

Government-sponsored R&D programmes are now a common feature of the science and technology policy landscape. Policy makers are also coming to understand that systematic evaluations of these programmes can also aid future policy formulation and implementation. Much attention has therefore focused on how best to conduct such evaluations, and one item under this heading concerns the choice of who should conduct them. Often this has been expressed in terms of a choice between academics or consultants, and it is this topic which constitutes the focus of this presentation. In particular, it shall be argued that much of this debate is far too simplistic in nature, depending in large part on crude caricatures of the nature and roles of academics and consultants, and on a weak appreciation of the factors influencing the choice of evaluators for different types of R&D programmes.

Exhibit 1 depicts some commonly accepted characteristics of academics and consultants respectively. Some of them deserve little elaboration. For example, academics are usually employed in the public sector, consultants in the private. Similarly, academics are normally expected to charge less for their services than consultants for an equivalent period of time, and this in part affects their relative availability. It is often presumed that consultants can only be afforded in short bursts, whereas academics can be utilised on a longer-term basis, though a corollary

of this is that the commitment of consultants to an evaluation can often be more focused and complete than that of academics with rival demands on their time during extended evaluation periods.

On a pejorative level, academics are often thought of as more profound and less analytically shallow than consultants, and as more objective and less self-serving, i.e. less likely to produce reports designed only to curry favour with their paymasters. To balance things out, however, academics are interest-driven to the extent that they can sometimes be diverted from the evaluation task in hand by an interesting line of intellectual enquiry - interesting, that is, to themselves, but not always to an expectant policy community hoping to learn something useful from an evaluation. Consultants are more problem-driven and less likely to make tangential excursions.

EXHIBIT 1

COMMONLY ACCEPTED CHARACTERISTICS OF ACADEMICS AND CONSULTANTS

ACADEMICS	CONSULTANTS
PUBLIC	PRIVATE
CHEAP	EXPENSIVE
LONG-TERM AVAILABILITY	SHORT-TERM AVAILABILITY
PARTIAL COMMITMENT	TOTAL COMMITMENT
ANALYTICALLY 'DEEP'	ANALYTICALLY 'SHALLOW'
OBJECTIVE	SELF-SERVING
INTEREST-DRIVEN	PROBLEM-DRIVEN

By now it should be apparent that the above generalisations offer only crude caricatures of both academics and consultants. Numerous counter examples and qualifiers spring to mind. Before examining why this is so, however, it is useful to describe one other commonly held perception vis-à-vis the tasks expected of academic and consultancy organisations. Exhibit 2 makes a distinction between policy analysis and technical assessment tasks related to the evaluation of R&D programmes, and between the types of institutions expected to carry them out. There are three broad categories:

- academics in science and technology departments undertaking technical assessment as part of the normal peer review process;
- consultants resident in Management Consultancies performing policy analyses;

- technical experts in a variety of private and public sector settings called upon to assess technological developments.

This categorisation scheme is useful not because it provides an adequate means of describing the current situation. It is in fact a very inadequate descriptive scheme. But it does allow us to see that the roots of this inadequacy lie in current developments which are blurring both task and institutional boundaries. For example, within evaluations based on technical peer review processes it is not unheard of for the experts involved to comment not only on scientific and technical excellence, but also on the efficiency with which a programme has been conducted, and on the appropriateness of the initiative as a whole. Thus there is a blurring of task boundaries. Equally, it is also possible for policy analysts in, for example, Science Policy Units, to be involved in considerations of technical merit. Familiarisation with technical developments in certain areas, acquired during the course of extended evaluations, can sometimes allow policy analysts to make positive contributions to technical assessments.

EXHIBIT 2

EVALUATION TASKS AND INSTITUTIONS

COMMONLY ACCEPTED PERCEPTION

	TASKS	
	POLICY ANALYSIS	TECHNICAL ASSESSMENT
ACADEMICS	Science Policy Units	Peer Review Process
INSTITUTIONS		
CONSULTANTS	Management Consultancy	Technical Experts

FACTORS AFFECTING THIS SITUATION

- THE BLURRING OF TASK BOUNDARIES
- THE BLURRING OF INSTITUTIONAL BOUNDARIES
- DIFFERENCES BETWEEN PROGRAMMES
- DIFFERENCES BETWEEN EVALUATIONS

There is also a blurring of institutional boundaries. It is fairly common these days for academics to spend some of their time on their usual pursuits and the remainder acting in a consultancy capacity, often under the banner of small consultancy firms. In turn, consultants occasionally act as Visiting Fellows or Professors at academic establishments. Furthermore, as we shall see, a number of evaluations utilise teams of both academics and consultants.

The blurring of task and institutional boundaries complicates a simple choice between academics and consultants in the evaluation of R&D programmes. There are also crucial differences between programmes and between types of evaluations which further complicate the issue. Exhibit 3 lists some of the key variables. With regard to the nature of the programmes being conducted, the type of R&D, the scale of the programmes and their technical scope are all important determinants in the choice. For example in large expensive programmes of industrial R&D the budget may be large enough to involve consultants, whereas in small programmes of academic research the percentage of the budget available for evaluation purposes may not be enough to attract consultants. Similarly, when the technical scope of a programme is very broad, the normal peer review mechanisms become complex and unwieldy, and analysts are often called in to collect and synthesise views on technical performance via interview- and questionnaire-based techniques.

EXHIBIT 3

CRITICAL VARIABLES IN THE CHOICE OF EVALUATORS

NATURE OF PROGRAMME	NATURE OF EVALUATION
TYPE OF R&D	TIMING OF EVALUATION
SCALE OF PROGRAMME	SCALE OF EVALUATION
TECHNICAL SCOPE	AIM OF EVALUATION
	STRUCTURE OF EVALUATION
	NATIONAL/INTERNATIONAL PERSPECTIVE
	PUBLIC/PRIVATE OUTPUTS

The nature of the evaluation required is a critical determinant in the choice of academics or consultants, and Exhibit 3 lists some important features which differentiate evaluations. For example, if an evaluation is scheduled to run alongside a five year evaluation programme - a so-called real-time evaluation - then cost considerations make it unlikely that consultants would be involved on a long-term basis, though there would be room for their partial involvement in an evaluation

which had a modular structure. The aim of an evaluation is also important. If the aim is to assess the quality of the outputs, there has to be a role for technical assessors; if it is to review the efficiency of implementation of a programme, policy analysts are more likely to contribute positively. Then there is the question of the type of evaluation outputs expected. Consultants are used to producing reports which are destined only for the eyes of sponsors, whilst academics normally expect to publish their findings in the open literature. This issue in itself can often be the deciding factor in the choice between academics or consultants.

SOME REAL EXAMPLES

The complex array of factors affecting the choice of evaluation teams can best be illustrated by some real examples. Exhibit 4 summarises five evaluations which have been conducted in recent years. All of them concern R&D programmes in Information Technology (IT). Brief details are provided below, together with the main lessons for the involvement of academics and/or consultants.

EXHIBIT 4

THE USE OF ACADEMICS AND CONSULTANTS IN FIVE EVALUATIONS

	ALVEY	IT4	FINSOFT	JCI	ESPRIT
TYPE OF R&D	IND+ACAD	IND (+ACAD)	ACAD (+IND)	ACAD	IND+ACAD
SCALE OF PROGRAMME	LARGE/5 YEARS	MEDIUM/3 YEARS	SMALL/3 YEARS	SMALL/5 YEARS	LARGE/5 YEARS
TECHNICAL SCOPE	BROAD	BROAD	NARROW	NARROW	BROAD
TIMING OF EVALUATION	REAL TIME	REAL TIME (X-SECTION)	EX POST	REAL TIME	EX POST
SCALE OF EVALUATION	< 0.5 %	< 0.5 %	< 1.0 %	< 2.5 %	< 0.1 %
AIM OF EVALUATION	APPROPRIATENESS EFFICIENCY EFFECTIVENESS (QUALITY)	APPROPRIATENESS EFFICIENCY EFFECTIVENESS QUALITY?	(APPROPRIATENESS) (EFFICIENCY) EFFECTIVENESS QUALITY	APPROPRIATENESS EFFICIENCY EFFECTIVENESS QUALITY	EFFICIENCY EFFECTIVENESS
STRUCTURE OF EVALUATION	MULTI-MODULAR	MULTI-MODULAR	BI-MODULAR	MULTI-MODULAR	BI-MODULAR
NATIONAL/INTERNATIONAL	NATIONAL	INTERNATIONAL	NATIONAL & INTERNATIONAL	NATIONAL & INTERNATIONAL	INTERNATIONAL
PUBLIC/PRIVATE	PUBLIC	PUBLIC	PUBLIC & PRIVATE	PUBLIC	PRIVATE
USE OF ACADEMICS/ CONSULTANTS	2 ACAD TEAMS TECHNICAL CONS POLICY CONS	CONS+ACAD ACADEMIC CONS+ACAD (PEERS?)	CONSULTANTS PEERS	2 ACAD TEAMS PEERS	CONSULTANT INTERNAL

The Alvey Programme[1]

The Alvey programme was a large programme of pre-competitive, collaborative R&D involving academics and industry. It ran from 1983 to 1988 and cost the UK Government over £200 million, with industry contributing a further £150 million, and it spanned a large number of technical disciplines, from Semiconductors to Software Engineering, and from Artificial Intelligence to Human Computer Interfaces. The sponsors of the programme wanted a real-time evaluation spanning the 5 year lifetime of the programme and beyond. It was to examine the appropriateness of the programme, the efficiency of its implementation, and its effectiveness in terms of goal fulfilment. The quality of the outputs was a secondary consideration, and the budget for the evaluation was set at less then 0.5% of the cost of the initiative to the UK Government. Evaluation outputs were primarily intended for circulation and publication in the open literature.

All these factors dictated an evaluation which consisted of a number of linked modules or packages conducted at different times over the life cycle of the programme. The need for continuous involvement of evaluators influenced the choice of academic teams over consultants. The diverse aims of the evaluation called for two academic teams with expertise in different, though complementary, areas, and the sheer breadth and scale of the programme, together with an ample budget, allowed the intermittent short-term use of consultants, both technical and policy-oriented.

Lessons from Alvey

- Single technical experts can provide credibility to an evaluation exercise, especially within the relevant R&D communities, but the utility of their input to policy analysis is limited compared to policy consultants.

- The credibility of academic policy analysts within the technical communities needs to be established and maintained via feedback and exposure to these audiences. This is more easily done within the context of long, real-time evaluations because of the opportunities for repeated exposure.

- Real-time evaluations allow experimentation and adjustments to the mix of academics and consultants.

---

[1] Evaluation conducted by the Science Policy Unit (SPRU) of the University of Sussex and the Programme of Policy Research in Engineering, Science and Technology (PREST), University of Manchester.

- Although independent peer review mechanisms and the use of technical experts are necessary to establish scientific excellence and quality levels, self-assessment techniques, i.e. those based on the achievement estimates of the participants themselves, are useful for comparing performance across programmes, especially those with a broad technical span.

- Collaboration between evaluation teams of academics and/or consultants can be synergistic, though sound steering group mechanisms have to be established to encourage convergence rather than divergence.

The Swedish IT4 Programme[2]

Running from 1987 to 1991, this programme was very similar to the Alvey programme in terms of its technical coverage and the nature of its participants, although there was less emphasis on collaboration with academics than in the UK example. The aims of the evaluation were also very similar to the aims of the Alvey evaluation. Again a real-time evaluation was requested, though the form it took was heavily influenced by the Swedish Government's decision to tap evaluation expertise outside the country. This made it impractical to have a full scale, continuous real-time evaluation. The Swedish decision to look outside its realms for an evaluation team was based partially on a perceived lack of relevant expertise within the country, and partially on a reluctance amongst certain Swedish academics to perform consultancy tasks over an extended period of time for academic remuneration rates.

The eventual IT4 evaluation package used the services of academic and consultancy teams. The first component of the evaluation used a joint academic-consultancy team in an intensive burst of field work during the first third of the programme, focusing in particular on the appropriateness of the programme in a Swedish context; on the efficiency of the programme start-up and early implementation phases; and on the establishment of programme and project aims for later use in assessing the effectiveness of the initiative. This initial period of field work was then followed by a questionnaire administered by the academic team, and a series of short visits over the middle period of the programme to receive updates on programme implementation and project progress. Finally, towards the end of the programme, the academic and consultancy teams combined again to carry out a final intensive burst of field work involving interviews and questionnaires, this time focusing much more on goal attainment and programme effectiveness.

---

[2] Evaluation conducted by Booz, Allen & Hamilton and SPRU in the first two phases, and by a spin-off consultancy, Technopolis Ltd. in the final phase.

Lessons from IT4

- Academics and consultants can jointly manage and conduct long-term evaluations.

- It is possible to structure evaluations into academic and consultancy phases.

- Whereas national evaluators are often better placed to comment on efficiency issues, non-nationals can be advantageous in terms of evaluating appropriateness and effectiveness.

- In some countries it is still difficult to find teams flexible enough to wear both academic and consultancy hats over the course of long evaluations.

The Finnish FINSOFT Programme[3]

In the late 1980s the Finns launched a small R&D programme in Software Engineering, to be conducted primarily in academic establishments and government Research Institutes, but undertaking applied work of relevance to industry. On its conclusion, an ex post evaluation was required to focus on two main aspects: the quality and scientific excellence of the research, and the effectiveness of the programme in terms of its industrial utility. A subsidiary aim was to comment on the appropriateness and efficiency of the programme.

Two separate evaluations were sponsored. The first involved a team of Finnish academics and consultants looking at industrial utility via interviews and questionnaires to members of industrial clubs set up to liaise with each of the research projects. The second strand involved the use of a mixed UK team of academics and consultants, policy analysts and technical assessors. Their role was twofold. The technical experts (two software academics and one industrial technical consultant) were to assess the scientific excellence of the work; the policy analysts (one consultant and two academics - all experienced in the evaluation of similar programmes in other countries) were to draw on their past evaluation experience and comment on programme administration and policy directions.

Lessons from FINSOFT

- Evaluation specialists and technical experts can combine effectively to produce ex post evaluations of technical quality and comparative performance.

---

[3] Evaluation conducted by Technopolis Ltd.

- Non-national peers have definite advantages in terms of quality appraisals. This is due to their independence and their often greater familiarity with scientific and technological developments in the world's leading research environments.

The UK Joint Council Initiative in Cognitive Studies and Human Computer Interaction[4]

In 1989 the UK established a small (£12 million) five year, academic inter-disciplinary research programme in cognitive studies and human-computer interaction. It was novel in a UK context primarily because it involved joint sponsorship by the three Research Councils covering science and engineering, social sciences, and the medical sciences respectively. The programme was designed to promote 'interstitial' work in the area, i.e. interdisciplinary work of interest to each Council but unlikely to receive funding from any one Council alone.

The novelty of the joint structure of the programme, together with a need to understand whether such a mechanism could be used in other interdisciplinary areas to cope with the problem of interstitial research, demanded a substantial evaluation effort. This was on top of the usual requirement to assess the scientific excellence of the academic outputs of a programme of this nature. With a limited programme budget, however, there was little scope for employing consultants. Two academic teams with considerable experience of working together were therefore employed to conduct a real-time evaluation, using their own experience and methodologies to evaluate administrative efficiency and effectiveness, and using national and international academic technical experts in a modified peer review procedure to assess scientific excellence. Of particular interest to the academic evaluation teams was the opportunity to develop methodologies adequate to the task of conducting real-time evaluation of a purely academic research programme. Earlier attempts at real-time evaluation had focused almost exclusively on R&D programmes much nearer the applied end of the R&D spectrum.

Lessons from ICI

- There is no or little scope for the long-term involvement of consultants in evaluations of small programmes.

- There is little incentive for academics to be involved in the long-term evaluation of small programmes unless there are elements of intrinsic interest or scope for experimentation.

---

[4] Evaluation conducted by SPRU and PREST.

The EC ESPRIT Programme[5]

Similar in nature to the Alvey and IT4 programmes, though much larger and more grandiose in conception, the EC ESPRIT programme has been the subject of a number of evaluations. One of these was initiated in 1960 by the European Court of Auditors. This body undertakes financial audits of the different activities carried out by the EC, and one of its traditional concerns has been the legality of processes and procedures. Recently, however, it has started to look at the value for money associated with EC programmes. For EC programmes with tangible outputs and achievements this presents few methodological difficulties for the auditors, but R&D programmes do not fall into this category. The traditional tools of the auditors were found lacking when it came to carrying out evaluations of R&D programmes. The solution, therefore, was to constitute a standard in-house team of auditors, but to draw also on the services of an external consultant to advise on the design of the ESPRIT evaluation and to act as an external monitor and critic of progress.

Lessons from ESPRIT

- There is a valid role for consultants with relevant experience in the design of evaluation and as monitors and critics of their progress.

General Lessons

There is no such thing as an easy choice to be made between academics or consultants in the evaluation of R&D programmes. There are, however, some general lessons which the commissioners of evaluations would do well to heed, and these can be summarised quite succinctly.

Know your programme

There is an overwhelming need to understand the nature of the programmes to be evaluated. The abilities and experience of evaluators have to be carefully matched to the type of programme they are expected to evaluate.

Know your evaluation needs and constraints

Be clear about the focus of the evaluation. An emphasis on determining scientific excellence demands a different type of evaluation from one with an emphasis on administrative efficiency.

---

[5] Evaluation advised by K. Guy of SPRU and Technopolis Ltd.

Understand the strategic as well as the tactical needs for evaluation
Often there is a statutory obligation to conduct evaluations, but the mere fulfilment of this tactical need is very rarely desirable in itself. The strategic need for evaluations lies in their potential to influence future policy formulation and implementation, and the quality of the evaluators and the evaluations they deliver is thus of vital importance.

Be wary of stereotyping academics and consultants
It is dangerous to perpetuate myths in a changing world. Tasks and institutional boundaries are blurring and the roles of academics and consultants are becoming harder to differentiate in any clear cut fashion. Often evaluations are best conducted not solely by one variety or the other, but by collaborative teams or the same teams structuring evaluations in academic and consultancy phases.

Look for flexibility in evaluation teams and approaches
Evaluation is still nearer a craft than either an art or a science. There are few mechanistic procedures which can be adopted to produce magic answers, and still only very few professional evaluators. By all means look for signs of competence when selecting evaluators, but be careful not to mistake exploration for ineptness. New approaches are necessary if the practice of evaluation is to improve.

Tore Olsen

A Comment on Academics and Consultants in the Evaluation of R&D Programmes

I have been wondering why I was asked to comment on Ken Guy's paper - being a physicist turned administrator.

I choose to think that the reason is that I have been involved in the commissioning of several evaluations in the last ten years. I will use this experience in my comments, and I will do this by using one particular example.

The example is the SPRU report: "Government Support for Industrial Research in Norway" from 1981. SPRU, the Science Policy Research Unit at the University of Sussex, was asked by the Thulin Commission, an ad hoc committee set up by the Government, to evaluate the extent, the organization, and the efficiency of publicly-supported and industrially-oriented R&D.

Hence it was a *policy analysis* on which the Commission should base its recommendations. It was, I think, one of the first R&D policy analyses performed in Norway and definitely the first where a foreign institution was involved. The Thulin Commission explicitly wanted an "outsider's" assessment.

But why did we pick SPRU and not one of the consultancy firms that we had on our list? This was seriously considered and the Commission was in contact with a well-known US-based consultancy firm.

If I look at Dr. Guy's list I can only find two pairs of contrasts that played a part in the decision process: "cheap - expensive" and "analytically deep" - "analytically shallow", and cost was not the dominating factor.

COMMONLY ACCEPTED CHARACTERISTICS OF ACADEMICS AND CONSULTANTS

ACADEMICS	CONSULTANTS
PUBLIC	PRIVATE
CHEAP	EXPENSIVE
LONG-TERM AVAILABILITY	SHORT-TERM AVAILABILITY
PARTIAL COMMITMENT	TOTAL COMMITMENT
ANALYTICALLY 'DEEP'	ANALYTICALLY 'SHALLOW'
OBJECTIVE	SELF-SERVING
INTEREST-DRIVEN	PROBLEM-DRIVEN

It was essential to us that the evaluators had a knowledge of the subject matter - R&D policy - and hence had the knowledge to perform a "deep" analysis and that they came from an institution where this was done on a regular basis. Let me allow myself a small metaphor. To do an evaluation you need a good tool kit - theory, methods, etc. - and experience in using them. Although consultants have more experience in handling these tools, they often lack knowledge of the properties of the material on which the tools are to be applied, in this case R&D. Academic institutions will be better on this point. I am a little hesitant to use management consultants that apply their skills on banks one day and research institutions the next.

The moral of the story is that an understanding of the subject matter, the traditions, the art of the trade is important.

The Thulin Commission and the SPRU report have had a significant impact on the Norwegian R&D system. Many of the recommendations are now well-established policy.

Why this success?

The reason for the successful outcome is not to be found in details in the evaluation process nor in the methodology but in an overriding determination among all involved to reach the goal, which in the present case was improved industrial innovation. Hence a deep understanding of the psychology of the decision makers on all levels was essential should the system be changed. Both the evaluation process itself and the recommendations, therefore, must be tailored to the needs of the implementation phase.

Against this background the large numbers of interviews performed both by the Thulin Commission and the SPRU group served the additional purpose of involving decision makers in the process from the start. This helped the implementation of the recommendations at a later stage. Furthermore, it was important that the recommendations were fundamental and based on an easily understandable philosophy. The Thulin Commission's basic philosophy may be summed up as follows: There must be more strategic long-term planning at the top level, more delegation of research planning to the performing level, and more coordination at all levels.

May I draw attention to Dr. Guy's lessons:

LESSONS FOR COMMISSIONERS OF EVALUATIONS

KNOW YOUR PROGRAMME
KNOW YOUR EVALUATION NEEDS
UNDERSTAND THE STRATEGIC AS WELL AS THE TACTICAL NEEDS FOR EVALUATION
BE WARY OF STEREOTYPING ACADEMICS AND CONSULTANTS
LOOK FOR FLEXIBILITY IN EVALUATION TEAMS AND APPROACHES
PHONE THESE NUMBERS:

SPRU	TECHNOPOLIS
0273-678175	0273-601956

Finally, we *did* call one of these numbers!

Luigi Massimo

Evaluation of the R&D Programmes of the European Communities

Evaluation fulfils a number of different functions. Evaluation of public policies and in particular of research can be seen from two different points of view: control and management. An external independent control of the use of public funds is essential in a democratic society and is an important element for political decision makers. However, it would be wrong to see evaluation from a negative point of view implying control and sanction. Its essential function is to assist management at all levels from political decision makers down to the people charged with the daily execution of the programme under scrutiny. Therefore, evaluation has become an integral part of the R&D management process and should not be seen as an exceptional action to be taken when problems arise.

It is important to distinguish between programme evaluation and scientific peer review. Scientists have been accustomed since a long time to a scientific and technical analysis of R&D activities conducted by their peers (peer review). While this continues to be an indispensable element of the R&D process, evaluation of publicly funded research programmes is intended to go beyond scientific peer review in order to analyze these programmes as R&D operators (see R. Chabbal: *Organization of Research Evaluation in the Commission of the European Communities*. EUR 11545, 1988). It becomes therefore essential to assess, beside the individual research projects, the managing structure of the programme in order to analyze the particular contribution given by public national or international intervention.

Seen under this point of view evaluation is a continuous function which takes place during all phases of the programme. It is primarily an internal activity conducted at all different levels of programme management. However, at given intervals in time, it is important to analyze R&D programmes under a more general perspective different from the one of the specialized point of view of their managers.

Public funding of R&D programmes, even in the case of basic research, is normally justified by short- or longer-term goals which go beyond the pure increase of scientific knowledge. History has proven that economic prosperity and quality of life are in the long run strictly related with past R&D expenditure, even if the relations of cause and effect are not straight forward and cannot be easily schematized. It is therefore essential that the best utilization of public funds be

regularly assessed from a point of view which cannot be limited to one of pure science and technology.

Evaluation by external experts is then the occasion to bring into the scientific chain of thought different points of view ranging from those of different but related scientific disciplines to the those of economists and management specialists. External evaluations conducted by independent people, beside fulfilling the function of democratic control, therefore also function by bringing in new schemes of thought.

We can therefore distinguish the following phases of evaluation:

- a general *ex-ante* definition of priorities, objectives and milestones,
- a continuous day-to-day evaluation which is part of the normal management functions
- an external independent evaluation which can take place either at the end of the programme (*ex-post*) or during the course of its execution (*mid-term*).

Ex-ante evaluation

The function of an ex-ante evaluation is to define as clearly as possible the objectives of the programme and plan its development as a function of time. A particular problem is posed by the definition and further interpretation of the objectives of R&D programmes.

In the past, these have often been very general, e.g. "to contribute to better knowledge of the marine environment" and "to encourage the development of new technologies for ... marine resources". However there is now a greater awareness among decision makers and programme managers that the objectives should be written in verifiable form, and so they are tending recently to be at once more specific and much longer and more detailed.

However, one should be aware of the need for objectives to respond flexibly to changed external circumstances, and that unexpected spin-offs may be so important that they can make the original targets almost irrelevant. An often-quoted example is the voyage of Columbus which failed dismally to meet its original objective and yet changed the course of history.

The objectives of a programme can be of two types, to solve a particular problem, or to cause particular things to happen. Both can in principle be stated in verifiable form. A famous example of the first type is Kennedy's goal of putting an American on the moon and returning him safely to earth before 1970. The latter might be exemplified by the requirement that European industry fund further development of the ideas contained in the projects with twice the money spent by the Commission. This would allow for the possibility that some projects would fail.

The writing of clear objectives is done not only to facilitate the task of the external evaluators, but even more to provide discipline for the programme managers, who thereby state what they intend their programme to achieve. It also provides appropriate signals to the programme participants and assists in the development of their plan of activity. It is thus a fundamental part of the management of a research programme.

The programme managers are asked to consider the current situation, and how they would like this to be changed and improved in, say, five (or ten) years time as a result of the implementation of their programme. There should be a demonstrable causal connection between the work undertaken under the programme, which is additional to what would otherwise have taken place, and the results intended. Whenever reasonably feasible, objectives should be expressed in a quantitative form and the means of testing them should be specified.

A good example of testable objectives is afforded by the BRIDGE programme in biotechnology. This includes a requirement for transnationality, to be expressed in multi-nationally coauthored papers, or ones with acknowledgements to other contract partners for the provision of materials and/or methods. Another requirement is for direct industrial involvement in at least one-fifth of the projects, either during implementation or afterwards.

The check of the fulfilment of objectives may require the collection of important amounts of information and is a non-trivial exercise. The evaluators may well feel constrained to make a selection among the evaluation criteria if they cannot check them all. In any event, it would not be reasonable to expect a programme to achieve every single one of its objectives, and some order of priority needs to be established. The check of the fulfilment of individual objectives will help the evaluators to reach a judgement on the success of the programme as a whole, but cannot replace this judgement.

Beside this definition of verifiable objectives, ex-ante evaluation is intended to plan the programme development as a function of time setting up the relevant milestones.

Internal evaluation

This function cannot be easily distinguished from the normal management of the programme. It is conducted by the programme managers with the help of their advisory committees and includes a peer review both for the selection of new proposals and for the analysis of terminated projects.

Internal evaluations should also put together all information and data needed for subsequent external evaluations. It is useful to make sure that such information is collected from the beginning of each programme. This should include the programme decisions, calls for proposals, selection criteria, list of proposals retained

and rejected, progress and final reports of each contract, published articles, patents, seminars, conferences, opinions of the advisory committees, etc. It is however very difficult to convince a busy programme manager to devote time to the preparation of an evaluation due to take place three or four years later. The best way to proceed is to make sure that the files and databases which have to be kept for the normal administration of research contracts also include the information needed for evaluation.

Timing of external independent evaluation

For the R&D programmes of the European Communities external independent evaluation has become a necessary process which is officially required whenever a programme has to be extended or modified. This has the advantage of eliminating discussions on the need for evaluations, but it implies a constant control of the quality of these exercises in order to avoid them becoming simply a bureaucratic hurdle.

Evaluations are required when decisions have to be taken about programme continuation, termination or re-orientation. However, it is a truth universally acknowledged that evaluations are always started too early and evaluation reports always come too late.

A good evaluation should be started when results are available or, even better, when scientific results have produced all of their social and economic effects. On the other hand, evaluation reports are needed when decisions have to be taken. Very often these decisions are required when the programmes have been in existence only for a short time and no scientific results are yet available.

An evaluation report published after the relevant decisions have been taken is good for science historians but useless for managers.

Therefore, real ex-post evaluations are seldom conducted. The main evaluation work is centred on mid-term analyses assessing the available results and the management structure of programmes. Furthermore external independent evaluation should, as we have seen, introduce different points of view in the management of R&D programmes, and this has to be done at regular intervals. Ex-post evaluations come too late for this function. Since Community R&D programmes often cover several multi-annual cycles, it is frequently possible to conduct a mid-term evaluation of the current activities and an ex-post evaluation of the previous programme(s) at the same time.

Sometimes there is a problem when a large number of proposals for different R&D activities have to be submitted concurrently for political decisions. It is indeed difficult to conduct too many evaluations in parallel in order to have their reports available just on time for decisions. In this case the Commission has made use of older evaluation reports accompanied by an update.

Panels and consultants

An external independent evaluation can be conducted either by a specialized organization or by a panel of independent experts. Organizations specialized in R&D evaluation are still rare. Most consultants are specialized in various technical fields, management or marketing. All of these functions are needed for evaluations but are seldom brought together in the same organization. Moreover expertise in the particular field of research evaluation is often not available.

At the level of the European Communities it has been felt that the use of panels can give a better guarantee of independence and have a higher political impact. European evaluations have to be accepted by the representatives of the Member States, by the European Parliament and by the scientific community. The involvement in this process of well-known personalities from different countries can strongly help in this respect. Furthermore, consultants are seldom multinational and are often seen as executors of the wishes of their customers rather than independent judges. In this respect the situation is politically very different from the one of a national agency asking a contractor to organize an evaluation for its own use.

The use of panels also gives the possibility of putting together expertise in a number of different fields. Indeed experience has shown that the best evaluations are those conducted by the most heterogeneous panels. If the panel members are too specialized in the technical field under examination, the discussion tends to concentrate on narrow issues and technical details and neglects the more difficult analyses of the general impact of the programme. One should not forget that decision makers must also be able to use evaluations to set priorities between different fields. This is only possible if the evaluation panel, beside the specialists of the relevant technical field, also includes specialists of different technical domains.

Indeed people who have spent much of their lives in research tend to believe that their field always deserves the highest priority, and only the inclusion of people with experience in other fields of research can guarantee the necessary objectivity.

Users of research results should be included, and particularly industrialists, whenever relevant. Expertise in science policy, management and economics is also needed.

The choice of evaluators

The choice of panel members is the most delicate part of an evaluation, influencing both its value and its credibility.

The independence of the evaluators is an important element if evaluations are to be used in the democratic decision-making process. Therefore panel members should not directly benefit from the programme and should at the same time be seen to represent different points of view in controversial cases (e.g. industry versus

environment). They must be sufficiently eminent to make the evaluation report credible.

A reasonable balance of nationalities must be obtained but one should avoid having a bureaucratic group of official national representatives. It is in any case impossible to include all Member States since an efficient panel cannot contain more than 7 or 8 members. Experts from outside the Community often add an important contribution, particularly for those programmes that have involved the quasi-totality of the scientific community of the Member States. However the inclusion of members from the USA or other distant parts of the world must be balanced against the problems posed by the long journeys, jet lag, costs, etc.

The method chosen by the Commission for the choice of panel members consists of the following steps:

- Drafting a list of profiles defining the types of expertise required and background sought (e.g. economist from industry specialized in raw material problems);
- Collecting a large number of names corresponding to these profiles. Suggestions are sought from many different sources including the programme managers, their management or advisory committee members, other Commission officials, and the database or other contacts of the evaluation unit;
- Checking independence (see below);
- Selecting a "short" list of possible panel members taking into account expertise, professional affiliation and a reasonable balance of nationalities. This list is formally submitted to the Director General, who may add additional names, or delete some.
- Inviting people on the list to serve on the panel. Very often the panel chairman is selected first and the other members are chosen with his help.

This selection process takes a long time. High-level experts, especially from industry, are not readily available and sometimes a short list of 25-30 names is needed in order to arrive at a panel of 6 or 8 experts.

Every time a proposed member declines to participate it is necessary to re-assess the balance of expertise, affiliation and nationality and contact other potential members. Some experts ask for documentation, analyze it and then declare that they have no time to participate so that more than one month is lost on a single refusal.

Based on an examination of six recent evaluations, the average time needed from the decision to start the procedure to the first panel meeting was 9 months with a minimum of 6 months and a maximum of 16.

The concept "independence" is also rather vaguely defined. It is almost impossible to find Europeans who have never benefitted in some way from the activities of the EC. The normal check consists in ensuring that they have not received contracts from the programme to be evaluated nor have participated in one of its committees. This check is not always easy. In the Medical Research programme, for example, approximately 4000 teams of researchers have been involved and some of the people who were originally proposed as independent had later to be excluded because they had participated in the research. Experts in the field covered by the programme are seldom totally independent even if they did not participate in its contracts. However, by involving people with different backgrounds, the panel as a whole can be more independent than each of its individual members.

Programme managers are allowed (within reasonable limits) to refuse specific persons they feel would be unduly biased against their programme and therefore lack independence.

The involvement of the programme managers and the members of their advisory committees in the selection process for the panel gives them more confidence in the evaluation process.

During the evaluations, it is a common experience that the panel members tend to develop a feeling of responsibility toward the programme they are evaluating. We have even found that, after some years, a few individuals who were originally independent have been retained to assist with the programme and they can no longer be used for subsequent independent evaluations. This does not mean that the original evaluation was not objective. Moreover, the evaluators have fulfilled their main role by introducing new ideas and different points of view into the management process.

External support

The use of panels of experts does not exclude the employment of external consultants. Indeed high-level experts are usually very busy and cannot devote a high percentage of their time to an evaluation. A considerable amount of the work needed for an evaluation requires specialized analyses of the programme both from the scientific and from the economic and sociological point of view. Besides scientific output, it is usually necessary to measure the impact of the programme on scientific structures and cooperation and its actual or potential effect on the European economy, industrial competitiveness, the environment, the quality of life, etc. The collection and analysis of these data require techniques only available through some specialized contractors. Therefore all preparatory work, such as collecting data, conducting interviews, mailing questionnaires, bibliometric studies, detailed technical or economic analyses, will have to be conducted by specialized

contractors. With questionnaires, it is particularly important that replies be treated confidentially by an organisation separate from the Commission so that the results are only made available in an aggregate form. Whenever possible the choice and terms of reference of these contractors should be made in cooperation with the panel in order to be sure that the results of these studies are fully accepted by and integrated in the work of the evaluators.

However this is not always possible because sometimes the work of the contractors requires many months. This creates a conflict between the importance of having the study conducted under the supervision of the panel and the need to start the work in advance in order to have the results available when the panel needs them. In some cases, particularly when the study was large and particularly expensive (e.g. a big programme of interviews) this problem has been solved by seeking tenders well in advance so as to be able to respond rapidly to the needs of the panel.

In some cases it is important to compare the situation before and after the programme so that the study has to be conducted twice. The first study has then to be conducted when the programme is starting, long before the evaluation, and only the second phase of the study can be supervised by the panel.

Terms of reference

In setting up an evaluation it is important that the task of the evaluators is clearly specified. This is usually done in the terms of reference which are part of the contract made with the members of evaluation panels.

For EC research programmes, some general guidelines were drafted in 1986 (Official Journal of the European Communities C 14 of 20.1.87). These general terms of reference state the need to assess both the scientific value and achievements of the Community R&D programmes and the added value resulting from their implementation at the European level. For programmes financed with Community funds it is not only necessary to show that they are technically and scientifically sound and properly administered, but also that Community action was justified and has resulted in some added value which could not have been obtained at the private or national level.

The EC terms of reference state that evaluations will cover the following :

- the scientific and technical achievements of the programme or activity taking into account its original objectives and milestones, and whenever relevant of changed circumstances,
- the quality and practical relevance of the results including (whenever relevant) commercial development and exploitation, and possible spin-offs,

- the effectiveness of management and of the use of resources,
- the programme's or activity's contribution to the development of Community policies and to the social and economic development of the Community,
- the benefits resulting from the implementation of the programme or activity at Community level (Community added value).

The first point (scientific and technical achievements) is usually dealt with in the internal evaluations or peer reviews conducted regularly by the programme management and their advisory committees. A programme evaluation conducted by a panel of external experts has the task of assessing the general impact of the programme and its rationale. It cannot analyze in depth every single project of which the programme is composed. Furthermore such work would require detailed expertise in all fields covered by the programme, which is usually not available in an external evaluation panel. A group of experts capable of analyzing the Community added value of the programme and the quality of management is anyway ill-suited for such a detailed task.

It is therefore essential that evaluators be able to base themselves on more detailed work conducted by other experts on each project during internal evaluations, and check only that this exercise has been done fairly and competently. Besides these project analyses, general output indicators (e.g. involving bibliometrics) can be used to complete the scientific picture of the programme. These analyses based on other evaluations have sometimes been called "meta-evaluation" (B. Bobe and H. Viala: *A Decade of R&D Evaluation at the Commission of the European Communities*, EUR 13097, 1990).

The general terms of reference we have just listed must be specified taking into account the characteristics of each specific research programme. Some will be aimed at helping industry and increasing its competitiveness, others will deal with environment, health and quality of life, while some will have more basic research goals. All of these specificities are of course detailed in the original decision that set up the programme together with its verifiable objectives and evaluation criteria.

On the basis of these original objectives a detailed mandate is then specified in which the terms of reference listed above are expanded into a number of questions suited to the specific goals of each programme. For example, in the case of the aeronautics programme the panel was asked to consider the following specific additional points :

- the contribution of such research to the technological competitiveness of the European aeronautical industry;
- the benefits accruing to technological areas other than aeronautics ;
- the added value of dedicated research in this area.

In other cases a much more detailed list was prepared. However it is important not to circumscribe the panel too tightly, partly because it could limit their independence, and partly because of the amount of time at their disposal. It may be helpful if the programme managers agree with the evaluation unit on a more detailed list of points to examine which can then be passed to the panel to guide them but not for them to follow slavishly. These detailed points are usually discussed with the chairman of the evaluation panel during the preparatory phase of the evaluation.

Considering that evaluations are not organized for historic purposes, but in order to improve future activities, their mandate always includes a requirement to give recommendations for the future continuation, alteration or termination of the programme or activity, for its management and for the use of research results either directly or through technology transfer.

In practice the question of continuation or termination of a whole R&D programme is seldom discussed by an evaluation panel. Panels have never considered whether to stop research on energy or on the environment entirely, but some parts of programmes have been stopped or re-directed following the recommendations of an evaluation.

Relations between evaluators and programme managers

Even if the main customers of the evaluation are the decision makers, its recommendations will have to be implemented by the programme managers who are also one of the main sources of information for the panel. Therefore a situation of conflict between evaluators and programme managers cannot lead to a good evaluation.

The fact that evaluation has become a necessary process in the management of EC research programmes has strongly reduced these conflicts because these exercises are not felt to have an exceptional or punitive character. Furthermore the situation can be improved by involving the programme managers in the various preparation phases of the evaluation, asking their opinion in the selection of the panel members and in the conduct of the supporting studies.

Sometimes scientists resent being evaluated by people who are not deeply specialized in their scientific field. They are accustomed to peer review and it must be clearly explained that the goals of a general impact evaluation of a programme are quite different from those of a scientific peer review.

During the evaluation there should be continuous contact between programme managers and evaluators. The programme managers must initially provide the panel with the necessary information on the programme, its goals and historical development, and its management structure and achievements. The results of the internal evaluations conducted by the programme managers must also be transmitted

to the panel. The panel must subsequently maintain a dialogue with the programme managers and keep them informed of their findings so that they can be taken into consideration in real time.

The panel must also make contact with the persons charged with the administrative aspects of programme management (e.g. the contract department) in order to avoid proposing administrative improvements which are too difficult to implement.

Individual meetings on a one to one basis between panel members and programme managers have been found extremely valuable. They allow the members to learn about the separate sub-programmes in much more detail, and they are apt to yield information that would not be vouchsafed in the context of a more formal presentation.

It is usual that while a programme is being evaluated by the panel, its next phase is being planned by the programme managers who should be able to make use of the evaluation results as soon as they become available. Furthermore, before an evaluation report is released, the programme managers should be able to see it and transmit their comments to the evaluators. The final decision on the report belongs of course to the panel, but this procedure is intended to avoid misunderstandings or factual errors.

However, the need for continuous contact between evaluators and evaluated does not mean that the programme managers should be present at all panel meetings. In particular, interviews with contractors or users of research should be conducted confidentially in order to obtain better information. This means in practice that the programme managers, or a representative of them, should be present only when specifically requested by the panel.

A practice which several panels have found very useful is to invite the programme managers to suggest the names of people whom the panel could usefully interview. The panel should, however, not confine itself to seeing only these people, and must retain its right to interview, or take written evidence from, anyone who may be able to give relevant information - even if this is not in accordance with the views of the programme managers.

In practice not all evaluations can take place in perfect accordance with these ideal procedures, and the personalities of the programme managers and of the panel members can in some cases give rise to some tensions. It is the task of the evaluation unit and of the panel secretary to avoid as far as possible such tensions.

The role of the panel secretary

The members of the evaluation panels are usually high level experts in different fields who are not necessarily familiar with evaluation. Furthermore they change from one evaluation to the next and the experience gained in one exercise would be lost for the following ones.

To avoid these problems, in almost all EC evaluations a secretary has been provided for the panel by the Commission's evaluation unit. This secretary plays a key role in the conduct of an evaluation. He (or she) is naturally responsible for making the arrangements for panel meetings and for the presentation of papers. He also advises the panel on how they can set about their tasks, what supporting studies could be undertaken, and the people who should be called to meet the panel or individual members thereof. In performing this task, he relies on the experience of his own and other colleagues' research evaluations, and on the activities pursued by the Commission in the field of evaluation methodology.

This enables him, for example, to prepare draft specifications for external studies, and to advise on their likely cost and the suitability of particular contractors. He also briefs the panel as necessary on the context of each programme and which other services of the Commission may be involved with the definition of the programme or with the use of its results.

In turn, the lessons learned from an evaluation and in particular from its supporting studies enable the panel secretary to make an effective input to the development of evaluation methodology which in this way reflects the real needs of ongoing evaluations. For example, a major bibliometric study performed for the Commission on measurement of scientific cooperation and coauthorship (F. Narin and E.S. Whitlow: "Measurement of Scientific Cooperation and Coauthorship in CEC Related areas of Science, EUR 12900, May 1990) arose directly from the needs identified in a small bibliometric study in connection with the biotechnology research evaluation.

The task of the secretary is a delicate one. He should not influence or bias the independence of the panel while at the same time he should provide a methodological guide. Since most panel members are his seniors, the suggestions of the secretary have to be given with a certain degree of diplomacy.

The panel secretary also acts as the main interface between the panel and the programme managers. He transmits to them the panel's requests for information and then presents this to the panel in the form that they require. In order to have an amicable professional relationship with them, he needs to explain the panel's and the manager's viewpoints to the other. Much of the success of an evaluation depends upon his persuading the managers of the reasons underlying the panel's conclusions and recommendations, so that they too become convinced that this is the route to follow and in a sense adopt the panel's views as their own.

The role of the panel chairman and the conduct of evaluations

The evaluation is the collective work of the panel, but this cannot be done without the coordination of a chairman whose task is to guide the meetings and to be responsible for the planning of the work.

To be chairman of an evaluation panel is a demanding task and it must be ascertained that the chairman has sufficient time to devote to this activity.

An important task of the chairman consists of creating a good team spirit among the evaluators. Experience has shown that problems have been posed both by chairmen with very strong personalities who conducted evaluations as a one man show, as well as by chairmen who lacked the strength to guide the work of the panel. This has to be kept in mind when choosing the chairman: a good and well-known expert may prove to be a bad chairman.

As we have already said, the chairman of the panel is often chosen first in order to discuss with him the panel membership. One or two meetings of chairman, secretary and programme managers usually take place before an evaluation is formally started. In these meetings the chairman is familiarized with the programme and its objectives, and possible evaluation procedures are discussed. Any studies which need to be started in advance are identified in this preparatory phase, so that firm proposals can be put to the panel at their first meeting.

The evaluation unit and the panel secretary brief the chairman, and later the other panel members, about available methods and current practices.

Evaluations conducted up to now by the EC evaluation unit have required between 4 and 9 panel meetings. These meetings usually last two days. This reduces the number of travels and the fact of spending an evening together tends to improve the team spirit of the members. There has been one case (the BRITE programme evaluation) where panel members have only been able to meet during week-ends. This was of course a heavy burden for the members, the secretary and the persons to be interviewed, but created a team atmosphere unequalled by any other panel. Between meetings panel members conduct interviews or visit laboratories in various countries either alone or in small groups, often accompanied by the secretary.

For an evaluation to be accepted it is important that all interested parties and countries be in some way involved in the process. Thus meetings of the panel or a group of members with officials in most or all Member States are often considered necessary. Depending on the characteristics of the programme to be evaluated, the panel seeks evidence from potential users of the programme's results, representatives of industry, consumers, trade unions, local authorities, environmental groups, etc. Some evaluation panels have solicited outside parties to submit evidence by sending a notice to the appropriate scientific journals. The response has not been particularly strong, but some written evidence was obtained in that way.

Since the more work-intensive collection of data, and studies are normally conducted by external organizations, the panel must concentrate on the most important interviews with managers, scientists, users of research and Member State officials.

The chairman usually subdivides the work among panel members according to their background, affiliation and nationality. Each member often contributes some part of the report which has then to be assembled by the chairman with help of the secretary. The final evaluation meetings are usually devoted to this task.

Utilization and diffusion of evaluation results

The publication of the results of external independent evaluations of programmes funded with public money is an important aspect of the democratic administration process. Evaluation is an important, even if not the only, tool for decision making in the R&D field. It contributes to this process by presenting reliable data and high level opinions. Therefore, it is important that evaluation reports be made widely available to political decision makers, scientists, to users of research and to the general public.

This also means that these reports are addressed to very heterogeneous readers. In order to be effective they must be easily readable. A good executive summary must be available for busy politicians; technical detail, if needed, must be confined to appendices. Work carried out for the panel by consultants may also be included in such appendices. It must be kept in mind that the document is often "used as a reference to check certain points, but not read in total" (PREST: *The Impact and Utility of EC Research Programme Evaluation Reports*, EUR 13098, 1990). For EC evaluation reports, translations in all official Community languages are also required, at least for the executive summary.

The current practice of the Commission of the EC is to publish evaluation reports, without necessarily endorsing their content, which remains the sole responsibility of the panel members. In this way the Commission retains its freedom of action while at the same time providing the decision makers with the independent opinion of the evaluators.

In theory, the evaluation panel may decide that certain material of a confidential nature (e.g. industrial property information or personnel matters) should be restricted to a confidential appendix and be communicated privately to the Director General or other appropriate person. In practice, in more than 10 years of evaluation this provision has never been used.

The dissemination of evaluation results is not a simple matter. Publication of the reports does not give any guarantee that they reach the right audience. Presentations to the specialized press and to a large public of scientists on the occasion of particular scientific events has proved to be an interesting method of diffusion.

Articles about evaluation reports by science journalists have sometimes shocked the programme managers because of their very negative interpretation: we all know that good news is no news. Therefore it is inevitable that the focus of the reporters is centred on the negative remarks of the evaluators.

An example is given by the evaluation of biotechnology. The report published in September 1988 did of course contain some criticism but was generally favourable and well balanced. It was sent to a number of scientific journals and a few articles were published, some of which chose to pick up only the points of criticism. On 15 October 1988, the report of the "*New Scientist*", under the title "Europe's Biotechnology Blues" started with the words: "Biotechnologists working on two of the European Community's research initiatives have failed to score any significant achievements". Another report of the "*Biotechnology Bulletin*", referring to the Biotechnology Research for Industrial Development and Growth in Europe (BRIDGE) had the title "A BRIDGE TOO FAR?" and started saying: "Over-ambitious targets, poor funding, a shortage of staff and a lack of co-operation with industry have all conspired to ensure that the European Community's two major biotechnology research initiatives to date have produced few useful results".

Reports of this type appear sometimes and should not scare programme managers and evaluators from disseminating the evaluation results. Experience has shown that it is better to be criticized than to be ignored and articles like these have enormously increased the readership of some evaluation reports thus helping to spread information about the programmes.

Cost of evaluations

Evaluation costs ranging between 0.5 and 1% of the total cost of the programme (or even higher) are often quoted. However these costs usually include those internal evaluations, peer reviews and ex-ante assessments which the Commission of the EC classifies as a normal part of the activity of the programme managers. The figure officially given by the Commission for external independent evaluations only, i.e. the cost of panels and related support studies, is of 0.25%. This is however an average figure which varies from programme to programme. Since the cost of a panel is not dependent on the size of the programme to be evaluated, large programmes are relatively cheaper to evaluate. Indeed one could estimate that a minimum expenditure of 60,000 to 70,000 ecus is needed for the pure operation of a panel. The cost of support studies varies strongly according to their nature. While bibliometric studies or questionnaires can be relatively cheap, large programmes of interviews and economic impact studies can be very expensive and will have to be justified from case to case.

Impact of evaluations

The impact of evaluations is twofold: on the decision-making process and on programme management. The first is conveyed by the report and essentially by its executive summary, while the impact on programme management takes place during the whole evaluation process starting with the setting up of verifiable objectives.

A recent study (PREST: *The Impact and Utility of EC Research Programme Evaluation Reports*, EUR 13098, 1990) after having interviewed a large number of programme managers and decision makers concludes that "where there was favourable timing most recommendations appear to have been implemented in subsequent programme planning". However evaluation is not the only tool of decision making and it may be difficult to identify a single cause for any given research programme modification. Whenever the panel developed a constructive dialogue with the programme managers many suggestions for change emerged naturally and were spontaneously adopted by the programme.

The utility of evaluation for programme managers is proven by the fact that many evaluators have subsequently become regular advisors to the programme and that several evaluation support studies have been further extended at the cost of the programme managers.

Hanne Foss Hansen

Process Evaluation - Possibilities and Problems

What in fact is process evaluation?

To begin with I consider it important to clarify what is meant by process evaluation. To me there seems to be some uncertainty, or even confusion, using this concept in the debate about the evaluation of research.

The confusion is due to the fact that the term "process evaluation" is used for describing at least three very different things.

Firstly, the term process evaluation is used for evaluations organized as dialogue processes. That is evaluations where dialogue between the evaluators and the actors or organizations being evaluated is considered important. Secondly, it is used for evaluations carried through in the process of a research policy activity - some call it "on-the-way evaluation". An example is a midterm evaluation of a research programme. Thirdly, it is used for describing evaluations in which processes are being evaluated, that is evaluations in which organizational processes, e.g. communication processes, Ph.D. education processes, management processes, are central objects.

Of course a specific evaluation may be a process evaluation in all three meanings of the term. But it may also be a process evaluation in only one way. The point here is that evaluators, and research administrators, or research politicians, when designing or ordering an evaluation, must consider very carefully if and in which respect it should be a process evaluation.

In the following I will discuss the strengths and weaknesses of each of the three types of process evaluation.

Process evaluation defined as a dialogue evaluation process

In theories about policy analysis and evaluation research (e.g. Premfors, 1989), we often distinguish between evaluations aimed at control and evaluations aimed at learning. The same distinction can be made in respect to the evaluation of research.

Figure 1. Two Models of Evaluation

	Model of Control	Model of Learning
Aim	Controlling object	Facilitating learning
Organization	Examination resulting in marks	Dialogue process
Evaluators	Superior level in hierarchy evaluates	Self-evaluation or evaluation by consultants or researchers
Diffusion of results	Upwards in the the hierarchy	To all interested parties and above all to the actors being evaluated
Use	Basis for sanctioning and controlling	Basis for improving organizational effectiveness by organizational change

Some evaluations have the purpose of control, for example controlling if society gets value for money from a certain research policy investment, e.g. a research programme. Other evaluations have the purpose of learning, that is developing consciousness of strengths and weaknesses in a department, a discipline or, for example, a research programme in order to improve organizational effectiveness in the future.

If the purpose is control, the evaluation will be organized as an examination. On the contrary, if the purpose is learning, it is essential that the evaluation is not organized as a hierarchical examination, but instead organized as a self-evaluation or carried through by consultants or, for example researchers, within the field of sociology of science. The assumption is that improvement of organizational effectiveness through organizational change implies acceptance of problem definitions and solutions as well as participation of the evaluated persons and organizations.

Likewise, if the purpose is control, evaluation results are spread upwards in the hierarchy and used for sanctioning and decision-making, e.g. in respect to resources. On the other hand, if the purpose is learning, evaluation results must be spread to several actors developing a dialogue process aimed at organizational change. In the learning model it is important that evaluation processes are decoupled as much as possible from decision-making concerning resources.

In social science, we call models like these ideal models. In practice, evaluations often aim at both control and learning. However, evaluators should consider which motive is the dominant one. If the purpose of learning is meant seriously, it demands that the evaluation is designed as a dialogue process.

Process evaluation defined as evaluation carried out midterm in a process

In respect to significant research policy activities, as earlier mentioned e.g. research programmes, priority areas and the like, it may be fruitful to evaluate midterm (as for example suggested by Narud & Søgnen, 1990). Most often the planning of programmes and priority areas is characterized by considerable uncertainty, especially uncertainty concerning implementation. Typical problems are to predict: Whether it is possible to raise research proposals of sufficient quality within the area in question? Whether the grant committee will able to make fruitful priorities?, etc.

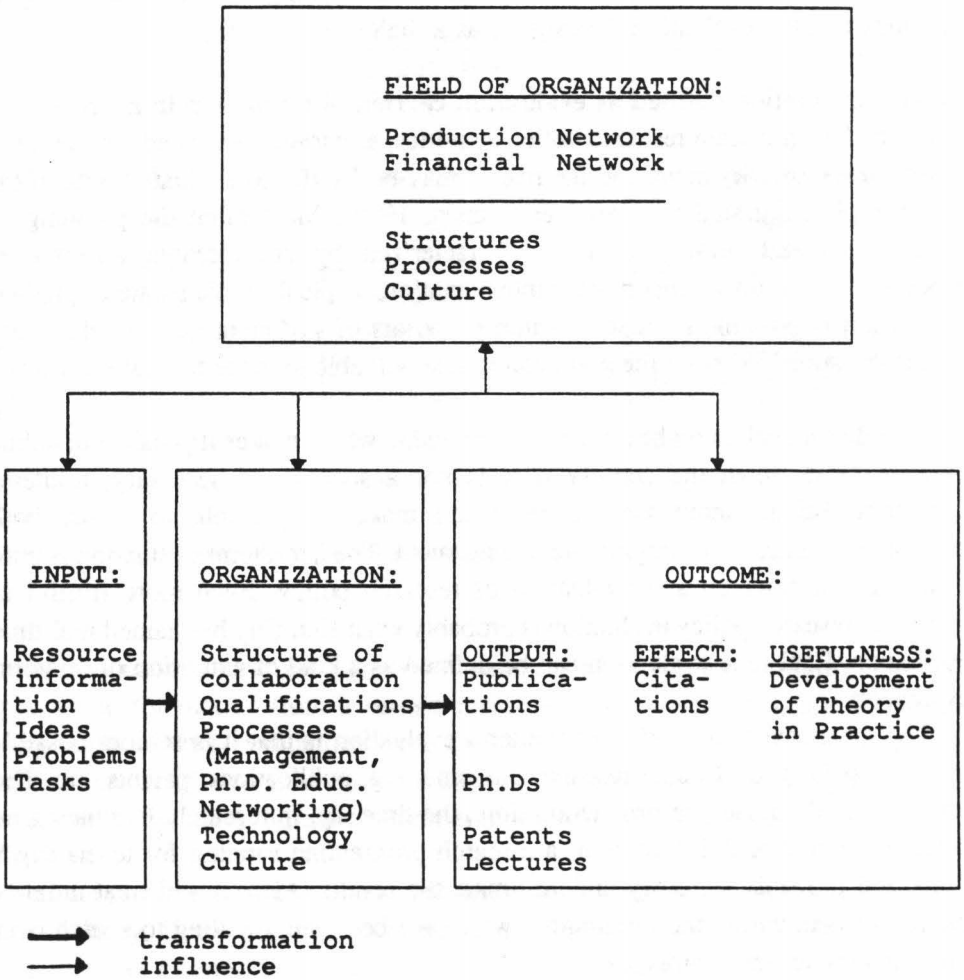
A midterm evaluation brings about knowledge which makes it possible to adjust or even close down the activity if it is not a success. If necessary, midterm evaluation brings about knowledge which makes it possible to adjust both programme content and programme management. Thus, midterm evaluation is both evaluation of research and evaluation of research policy. Even more fruitful in respect to research policy evaluation is probably what Ken Guy has named real-time evaluation. Real-time evaluation can be defined as a kind of extension of midterm evaluation.

The most severe limitation of midterm evaluation is that it is seldom possible (or at least fair) to evaluate research outcome, e.g. publications, patents, etc. The problem is, of course, the production time, the time lag, in research. Evaluating an outcome after 2 or 2 1/2 years in a research programme running for let us say 5 years will probably seriously underestimate the results. Also, it will treat unfairly the researchers within the programme who have been most willing to switch over towards new research problems.

Process evaluation defined as an element in organizational research evaluation

Figure 2 presents a very simple model of a research organization. Most likely you will think of a research department looking at the model, and that is okay. However, I would like to stress, that the model can be used for describing a cluster of departments or a discipline as well.

Figure 2. A Simple Model of Research Organization



According to the model a research organization is an open, resource-dependent organization, interacting with groups in the environment, e.g. with professional colleagues discussing the development of problems, methods and results, called the production network, and with groups financing research, called the financial network. It is characteristic that both environment and the organization itself influence production possibilities. The organization is characterized by structures, process-traditions and technology (methodology and equipment). And through these an outcome is produced.

Outcome can be divided into output, e.g. publications, patents, and Ph.D.'s.; effects, e.g. citations, and finally the usefulness of the knowledge produced in the further development of theory as well as the development of practice.

Within the research system, production possibilities can be very different, even between research organizations which appear very similar, e.g. two physics departments at different universities. Because of differences in production possibilities, direct, systematic comparison (for example through performance indicators) is often difficult (and we could add unfair).

What is the purpose of this very short presentation of this model?

First of all it can be used for classifying evaluation methods.

Peer review, in its classical form, is an evaluation method used for appraising the quality of the output of the research organization in question. Peers are good at evaluating scientific quality in respect to publications, persons, etc., and they are good at evaluating up-to-dateness in technology, that is in methodology and equipment. Also they are able to appraise if research problems are of scientific relevance, but not if they are of industrial relevance.

Bibliometric evaluation is an evaluation method used for appraising the quantity of output and the quantity of effect. Bibliometrics primarily uncovers productivity and visibility.

In other words, the classical evaluation methods can tell us whether research quality is good or bad, whether productivity is high or low, and whether the research production in the organization in question is noticed by other researchers.

To use a metaphor from medicine, both peer review and bibliometric evaluation methods have their strengths in the fact that they are very good at making symptoms visible, both symptoms concerning illness and symptoms concerning health in the research area in focus. The weakness is that these methods do not help us arrive at a diagnosis. In other words, they are not able to explain why research quality is good or bad, why productivity is high or low, or why research is noticed or not.

To make the diagnosis we need other methods, for example, to do an organizational research evaluation, that is, we have to analyze and explain

symptoms. To do this it is necessary to analyze processes as well as structures in the research organization and its surroundings.

In order to give some examples, organizational research evaluation makes diagnoses and suggests treatments around:

- collaboration structures;
- networking nationally and internationally;
- reward systems and incentive structures;
- financial structures and possibilities;
- organization of Ph.D. education;
- the quality of the research atmosphere;
- etc.

In organizational research, evaluation recommendations, among other things, can be deduced from the existing knowledge about excellent research organizations. From the sociology of science (for an introduction see Elzinga, 1986 and Foss Hansen, 1988) and from that part of psychology which is concerned about creativity (e.g. Amabile, 1983), we know that excellent research organizations are:

1. highly communicative organizations, internally as well as externally, nationally as well as internationally. In other words their production network, their networking with other researchers and research organizations, is very well developed;

2. characterized by variety in research profile, that is they are multifarious with both a profound and a broad competence;

3. characterized by freedom to choose research problems. Surveillance from peers or research politicians does not promote creativity;

4. characterized by stability towards scientific traditions (paradigms), at least for some time, and by ability to turn over the traditions to new generations of researchers. Consciousness in respect to tradition demands that other traditions are well known (this again stresses the aspect of being highly communicative).

Our knowledge about excellent research organizations can be summed up like this. Excellent research organizations are characterized by both:

- challenge and security,
- stability and change.

Some (for example Premfors, 1986) have used the term *structural instability* to summarize the complexity of this kind of organization.

Thus the strength of organizational research evaluation is that it is a knowledge-based and not "only" an experience-based evaluation. However, knowledge and research results within the field of sociology of science are in many ways uncertain or even in conflict. This of course is the weakness of organizational research evaluation. In fact, we need to do much more research about research organizations.

Recommendations for action

Concluding this discussion about definitions of process evaluation, I would like to put some recommendations for action to discussion. The message above has *not* been that peer review and bibliometrics are of no use in the evaluation of research. However, the message has been that it is important to recognize the limitations of these methods. Thus, my recommendations for action are as follows:

1. Peer review and bibliometrics are methods which primarily discover symptoms. Making a diagnosis and suggesting a treatment require the use of other kinds of methods, e.g. organizational research evaluation, including process evaluation.

2. In the evaluation of research we ought to be more experimental, for example by using different evaluation methods in the same evaluation.

3. Make experiments with interdisciplinary research evaluation, for example:
   - let peers, researchers within the sociology of science and organizational theory work together in a team,
   - make different evaluators evaluate the same object in order to make conflicting evaluation results visible,
   - make symptoms evaluation by using peer review and/or bibliometrics and, if problems are discovered, follow up by using organizational research evaluation to establish a diagnosis and suggest a treatment.

4. If you head for sustainable impact, organize evaluations as dialogue processes to secure organizational change.

5. If you wish to evaluate longer-lasting research programmes, priority areas, targeted programmes and the like, do it real-time or if you wish a cheaper solution, midterm. Thereby you secure both research evaluation and research policy evaluation. Also, midterm evaluation gives you a possibility to reorganize the activity.

Literature

Agersnap, Torben og Hanne Foss Hansen: *Forskningsorganisatorisk midtvejsevaluering af Det Bioteknologiske Forsknings- og Udviklingsprogram.* Økonomiske incitamenter som middel til udbygning af tværinstitutionelle forskningsnetværk. København: Forskningsdirektoratet, 1990.

Albæk, Erik: *Fra sandhed til information. Evalueringsforskning i USA - før og nu.* København: Akademisk Forlag, 1988.

Amabile, Teresa M: *The Social Psychology of Creativity.* New York: Springer, 1983.

Elzinga, Aant: *Kreativitet, Paradigmteori och Social Epistemologi - ett vetenskapsteoretiskt diskussionsinlägg om kreative forskningsmiljöer.* Göteborg: Institutionen för Veteskapsteori, Rapport 147, 1986.

Foss Hansen, Hanne: *Organisering og styring af forskning - en introduktion til forskning om forskning.* København: Nyt fra samfundsvidenskaberne, 1988.

Foss Hansen, Hanne: Programevaluering: Videnskab eller politik? Artikel i *Politica,* 23. årgang, nr. 1, 1991.

Narud, Hanne Marthe og Randi Søgnen: *Evalueringsoppleg for hovedinnsatsområdene. En drøfting.* Oslo: NAVF's utredningsinstitutt, Melding 1990:1. (English summary: Evaluation Plans for Norwegian Research Priority Areas. A Discussion)

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

Premfors, Rune: *Forskningsmiljön i Högskolan - En kunskapsöversikt.* Stockholm: Department of Political Science, rapport 36, 1986.

Premfors, Rune: *Policyanalys. Kunskap, praktik och etik i offentlig verksamhet.* Lund: Studentlitteratur, 1989.

Karl Erik Brofoss

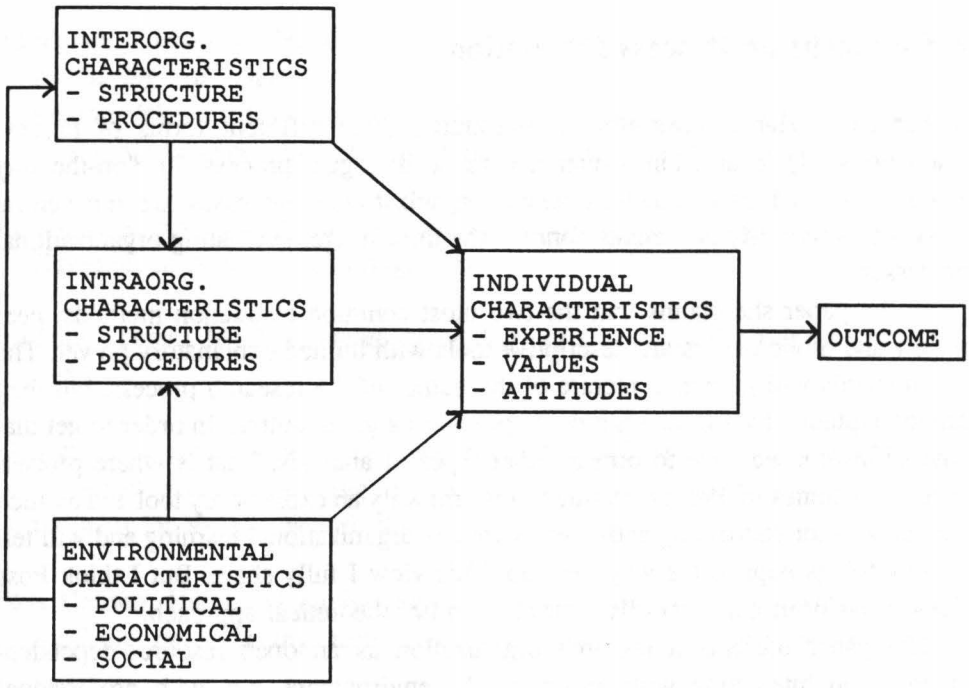
A Comment on Process Evaluation

In her paper Hanne Foss Hansen introduces three different forms of process evaluations: 1) evaluation organized as a dialogue process 2) "on-the-way evaluations" and 3) evaluations where organizational processes are the central objects of study. My comments concern the third usage: evaluating organizational processes.

In the paper she claims that the two most common evaluation methods, peer review and bibliometrics are descriptive tools with limited explanatory power. The two methods will give us a picture of the output of the research process, but they cannot explain why a research process produces a given output. In order to get that kind of insight we have to turn to other types of analysis. That is where process evaluation comes in. Process evaluation is primarily an explanatory tool and as such a basis both for controlling activities as well as organizational learning and will tell us why things happen the way they do. This view I fully share. But I think Foss Hansen has been unnecessarily restrictive in her theoretical approach.

The paper presents a research organization as an open resource-dependent organization interacting with groups in the environment, e.g. with professional colleagues and with groups financing research and is characterized by structures, processes and technology whereby an outcome is produced. By taking this approach Foss Hansen is virtually saying that process evaluation is nothing else than ordinary organizational analysis. And I agree. I think that process evaluation is just a subclass of a much more general analytical paradigm - implementation theory. Without doing violence to her model this can be shown in the Figure on the next page.

There are some differences between the two models. Firstly, in our terminology the field of organization in Foss Hansen's model will just be a part of a wider environmental concept being subdivided into political, economical and social characteristics. I fully agree that the characteristics pinpointed by Foss Hansen probably are the most important ones, but there is no reason, a priori, to restrict ourselves to those. The organizational structures in which process evaluation problems are embedded, will differ widely from case to case and so will their environments. Consequently I suggest that the concept of environment should be as wide as possible in order for us to capitalize the whole spectre of literature dealing with the interaction between an organization and its environments.



However, the main difference between the two models manifests itself in the way they deal with interaction between organizations. In Foss Hansen's model this aspect is partly dealt with as factors in the interaction between the research organization and its environment, and partly as factors within the organization itself. To my mind this hides the fact that as the research community becomes all the more diversified, interaction between different fields of science, and accordingly different value systems and research organizations, will steadily increase and become a very significant factor affecting the outcome of the research process. In order to illuminate this fact, I think it's necessary to include this interaction factor explicitly in the model.

Similarly I believe it necessary to explicitly include individual characteristics in the model. Elements such as experience, individual values and attitudes are essential in understanding a research organization and are of course a vital part also in Foss Hansen's model. But again, I believe that it's absolutely essential to draw the attention of the process evaluation practitioner to this fact by including individual characteristics separately in the model.

The scope and complexity of the evaluation problems at hand and the multiplicity of explanatory factors indicated above, suggest that there is no reason to believe there will be only one theoretical approach which will fit the bill. On the contrary there is every reason to believe that this will *not* be the case. A multiplicity or theories, approaches, frameworks and so forth should consequently be part and parcel of every process evaluator's tool kit.

Summing up: we can and should draw on the vast wealth of organization theory when we are doing process evaluations and not restrict ourselves to just one approach.

Some implications:

- Organization theory is not a unified theory. Consequently there is no unified approach to a process evaluation and there is no such thing as a standard process evaluation design.
- When we are doing a process evaluation we have to design a study tailored to the case at hand.
- Process evaluation demands methodological skills and should not be left to people doing self-evaluations.
- Since each study is unique, they often are resource-demanding and expensive.
- Process evaluation will also be more time-consuming.
- Given the time and resources necessary to do a proper process evaluation, you should think twice before you ask for a process evaluation.
- You should therefore have a very clear objective for the process evaluation - it should be part of a plan of action - otherwise you should not start a process evaluation.

Brit Denstad

A Comment on Process Evaluation

I have been asked to comment on the use of process evaluation in the work of research councils - specifically in the Norwegian Research Council for Applied Social Science (NORAS). The Council has existed for four and a half years and thus draws upon limited experience.

In total, NORAS has initiated external evaluations of four major research programs during these years - all conducted after the programs were concluded - and one "on-the-way" evaluation of the beginning and organization of a major ten-year long investment in research on management, organization, and administrative systems. Finally, NORAS has recently concluded an evaluation of three institutes for applied, regionally-oriented research. In addition, we have instituted a system of internal process evaluation - anchored in the program steering committees. I'll return to this.

The ambition of the evaluation activities, of course, is to develop knowledge and insights to make all parties in applied social research do a better job - whether in research, as organizers of research, or as "users" of research.

Evaluation studies within the field of *applied* research have to cover a complex process, focusing not only on questions of scientific quality, but also on questions relating to whether the problems raised are relevant to identified needs for knowledge, whether the results are effectively disseminated to identified target groups outside the research community, and whether the research process has been efficiently conducted.

Thus, it is important to be clear about *the more specific purpose* of evaluations, and to be emotionally, intellectually, and organizationally prepared to follow up results and recommendations if they are found sound and sensible.

Generally it is NORAS' strategy to be selective. Evaluations take time, effort, resources. We have to be fairly certain that it is worth the effort. We also want models and methods that can be fairly easily applied. And, we have developed a solid respect for the unintended consequences of evaluations - particularly, of course, when it comes to institute evaluations. Such unintended consequences may prove fruitful, but may also be destructive and out of tune with the quality and results of the evaluation study.

So much for general comments. I'll restrict the more specific comments to process evaluations directed at research programs and program organization - simply because NORAS relies heavily on this model of organization. In 1991 NORAS has organized sixteen research programs with program steering committees

in charge. These committees hold key positions in the steering system of the council.

On the program level NORAS has established a *system of internal evaluation of research programs* on an annual basis, but we also initiate *external, independent evaluations of selected programs* - focusing on initiation and the first phase.

I will stick to "process evaluations", and not comment on institute evaluations, nor on evaluations of finished programs and products.

Hanne Foss Hansen stresses the importance of being clear about "if and in which respects an evaluation should be a process evaluation", and whether the evaluation is primarily focused on learning or control.

This is a timely reminder.

However, our experience is also in line with her reminder that in practice there are blurred borders between categories.

That also goes for the distinction drawn between learning and control as a main objective. The purpose both of an internal process evaluation and an external, selective "on-the-way" evaluation is, of course, to learn - to establish a basis for adjustments and initiatives. But insights about how a program runs in relation to identified aims obviously also produce important knowledge to be used for control purposes, e.g., when it is found that institutes engaged in program research do not follow up on the obligation to offer professional guidance to the younger researchers engaged - and there prove to be deficiencies in the quality control system. It is the responsibility of the program steering committee to take action in a case like this. The point is that even though there is an analytical distinction between learning and control purposes, the distinction is not very helpful in practice.

I have mentioned earlier that evaluation studies in the field of *applied* social research have to cover a broad and complex subject matter, where evaluations of scientific quality alone only make up part of the picture.

One serious shortcoming we have faced in our evaluations so far is related to the lack of established, generally accepted *criteria for applied social research.* Criteria that can be related to the work of researchers as well as to the work of research institutes and the running and results of programs.

NORAS is now developing a set of criteria adjusted to the task of applied research. It is our ambition to establish standards that may serve as a common frame of reference for these types of evaluations. And we hope to develop criteria that may be useful both as a learning instrument for institutions and research councils, as well as for identifying areas where correction and control are necessary.

The critical question, of course, is whether, in our experience, these kinds of internal and external process evaluations function. The ambition of the system for internal program evaluation is to make the program steering committees

- formulate and set goals for the program
- evaluate status annually in relation to the goals - and so establish a basis for changing the course of direction, take initiatives of various kinds - and
- generate insights into the annual "evaluation" NORAS' *Board* carries out to establish whether the Council as a whole tackles its tasks in a satisfactory manner.

The internal process evaluation on a program level thus also serves as one of the mechanisms generated to make NORAS function as *one* institution.

Generally, we feel this system is beginning to function according to intentions. It also provides background material for the Board's choice of programs for external process evaluation. But we are concerned that this annual planning and evaluating process should not be allowed to disintegrate into an empty ritual.

Our very restricted experience with an "on-the-way" evaluation also gave useful results. The program on "Management, Organization, and Administrative Systems" benefitted through insights that resulted in a certain reorientation in the role of the steering committee and an increased weighting of strategic functions and a thematic concentration of the research at the program's Centre. The evaluation also established a basis for new research initiatives.

But process evaluation also involves a question of timing in relation to processes in the program. If the evaluation draws a line at a dynamic point in the program development, the results may lose their importance by the time the evaluation report is published. This underlines the importance of dialogue between the object of the evaluation and the person or team doing the evaluation study.

We will continue to do process evaluations of this kind, e.g., in areas where research raises particularly demanding problems (research on "modern crime"/"economic crime") or where we want to try out a "new" model of organization, and where the program is supposed to run for a fairly long time.

One limitation of process evaluations is often related to the fact that research results - the quality and importance of the products - cannot be satisfactorily covered. This task for evaluation has to be taken care of through other types of evaluations. But it is our experience that the learning involved for the organization is important for giving research better conditions in the future. The outcome, of course, first and foremost depends on the ability and willingness to draw actively on the insights generated in all parts of the system.

Terttu Luukkonen
Bertel Ståhle

Follow-up and Use of Evaluations

Introduction

By 'evaluation' we mean systematic assessments of scientific research that cover a wider area than individual scientists, specific research proposals or articles. In Norway, Sweden, Denmark, and Finland research councils and other funding agencies have adopted similar methods and procedures to evaluate such wider areas, for example, research fields, institutes and programmes. The model for evaluation in Scandinavia was developed by the Swedish Natural Science Research Council, and adopted by the other Nordic countries, with some modifications.

The gist of the method is the use of a panel of experts from foreign countries. The panel gets acquainted with the area under evaluation, first, by means of written information, and subsequently by site visits during which it discusses with scientists who work in the evaluated units. In order to obtain an analysis of such a wide area, the evaluation pays attention to its component parts, such as research groups, university departments, or smaller units in research institutes. Evaluation reports usually contain some general comments on the evaluated area and reports on the component units. They pay attention to the quality of work and problems in the organization of research.

In Sweden the Natural Science Research Council started the evaluations described here in 1976-77. Research councils in the other Nordic countries followed this example somewhat later in the 80s (Luukkonen & Ståhle 1990). There are two basic differences between the Swedish model and its Nordic applications. Firstly, while in Sweden the evaluations have concerned research activities funded by the research councils or other agencies which commission the evaluations, in the other Nordic countries the evaluations have assessed work funded also from other, but mostly public sources. This is an important distinction, since in the case of Norway, Denmark and Finland, scientists have consented to be evaluated by agencies other than those which fund their work. They recognize the importance of participation in evaluation for their future opportunities to get research grants. Secondly, in Sweden the Natural Science Research Council evaluates research areas systematically, and has so far completed over 60 evaluations. In the other Nordic countries the research councils carry out such evaluations on an *ad hoc* basis, and have completed a far smaller number of evaluations.

We have been involved in two studies of the uses and effects of Nordic evaluations. The first study was carried out in 1987-88 and was aimed at analyzing

the uses and role of evaluations in decision-making, in particular in research council organizations (see Luukkonen & Ståhle 1990). The second study is ongoing, and its purpose is to analyze the uses and impacts of evaluations from the point of view of the scientists and scholars who have been evaluated. It also examines the evaluation process critically. Both studies utilize written sources of information, but in particular interviews. In the first case, we interviewed key persons such as research council members who represented those who had commissioned evaluations, and secondly persons who had been responsible for their organization (N = 45). In the latter study, we interview scientists and scholars who have been evaluated (N = approximately 100).

Since the interviews with evaluated scientists are still ongoing (most of them have so far been carried out in Denmark and Finland), we will only make a few tentative and preliminary observations. Our paper draws mainly on the findings of the first study.

Science policy background of evaluations

The adoption of new and similar evaluation policies nearly simultaneously in four Nordic countries can be explained by tighter fiscal policies and rising research costs. There has been a need to legitimate an increasing or even level funding, especially when assigned to basic research.

Since the early 70s, governments in four Nordic countries have formulated explicit policies for science and technology, but within these policies investments in the development of new technology have been more easily accepted and justified than the funding of basic research. Most growth in research funds since the late 70s has been allocated to mission-oriented research and the development of technology.

The circumstances in the four Nordic countries vary. In Finland evaluations were started to provide arguments for more money for basic research. In Norway the great expectations raised by the discovery of oil (in the early 70s) induced an economic expansion and large increases in R&D budgets. However, difficulties in the exploitation of the oil fields and decreasing oil prizes a few years later created a situation in which all the ambitious programmes could not be accomplished and investments had to be submitted to a stricter screening.

In Denmark evaluations provided a means to legitimate a decrease in the number of university personnel and a dismissal of even tenured persons after the government had decided to decrease university funding as part of a savings programme in the public sector in the mid-eighties. In practice, this policy meant a reallocation of funds from institutional support for basic research to the funding of large research programmes.

In Sweden the strong tradition of planning in the public sector provides an answer for the upsurge of evaluation activities: evaluations became part of the

rationalistic ideology of state administration developed during the long era of social democratic government. The institutionalization of publicly financed research activities is highly developed in Sweden. The research councils have been given an explicit task by the Government to evaluate their activities.

An indication of the general science policy motives for evaluations was the fact that they were started at the request of science policy organs higher up in the hierarchy than the research councils which performed them in practice.

Evaluations, whose tool?

We have outlined above the general science policy background of evaluations. It is difficult to judge to what extent they have succeeded in securing research funds or preventing a less advantageous financial development for research than has in fact taken place. We heard a comment in Sweden that the evaluations had enhanced the position of the Natural Science Research Council and contributed to a sizeable increase in its resources. It may also be that, if a legitimation of research funding and the funding mechanisms are successful, there may be no change in funds or institutional arrangements.

One of the non-observable general impacts of evaluations is their role as a mechanism of quality control. Evaluations provide a legitimation for the funding system, but also affect the general work climate and, supposedly, enhance the quality and quantity of work.

We will pay more attention to the uses and effects of evaluations in the decision-making of research councils or universities and impacts on scientists and scholars.

The studies we have referred to pinpoint that evaluations are more useful for research councils or other funding agencies than for the scientists and scholars who have been evaluated. The latter are able to observe relatively few effects and benefit from evaluations to a lesser extent.

This is understandable given that the research councils or other funding agencies commission the evaluations and these are presumably tailored for their use. Besides, an average scientist or scholar is not in such a position that he could see the effects very well even in his own case. He does not know, for example, whether a decision to grant or not to grant him research money is based on the evaluation report.

Different uses of evaluations in decision-making

The research councils or other agencies use evaluations in their decision-making in several ways:

1. Evaluations provide supplementary information for project selection in the research councils. Even though the primary objective of evaluations is to assess performance in research fields, institutes or programmes, they usually comment on projects, departments and other small units within the evaluated field. If an evaluated scientist or a research group applies for money to a research council, the comments can be utilized. Applicants also utilize evaluation reports to support their case with the funding agency. Nevertheless, evaluations seldom play a crucial role in decisions on grant applications.

2. Decisions on the funding of instruments, deployment of people and modification of programmes are further examples of cases to which evaluations contribute. Even though research councils commission evaluations and are their primary users, other organizations such as universities or ministries have utilized them in budget plans or in the allocation of funds between university departments.

   In Denmark universities have been subjected to cutbacks in funds in the 80s. This has in practice involved even dismissal of permanent personnel. Evaluations of research fields by the research councils have been used to legitimate planned dismissals or to select the people to dismiss. Universities have also introduced their own evaluation mechanisms and set up minimum requirements for research performance to select those who will be subjected to dismissals. For example, the University of Copenhagen monitors the number of publications, and especially of publications in refereed journals, by its personnel and has used and will use such information to target dismissals. This procedure is more important for the management of university resources than the few evaluations of research fields that the research councils have so far commissioned and has far stronger impacts on the general work climate of the university.

3. Evaluations pinpoint problems in the organization of research such as insufficient collaboration, a lack of joint use of facilities, too much parallel work or gaps in research topics, inbreeding, inflexible career structures, and problems in funding arrangements. Often the problems had been known before, and the evaluations emphasized their importance. The authority of recognized foreign scientists, however, is needed for the acknowledgement of problems. Sometimes, evaluations help to formulate questions that one had not thought of before.

It is often difficult to act on the basis of such general comments. The remedies for problems would demand measures from several authorities or, sometimes, a radical change in attitudes. For example, inflexible career structures need measures from the universities, the Ministry of Education, and the research councils. In order to be able to cure inbreeding we might have to undermine the current practice in the Nordic countries that scientists pursue a career in one and the same university from their first degree until their retirement. This might involve an introduction of new rules which, for example, forbid the employment of scientists who have taken their PhD at the same university for a fixed period of time after the degree. Forced mobility, nevertheless, would cause problems for spouses and children and go against efforts to enhance women's career developments.

Factors which affect use and effects

Many of our interviewees disapprove that evaluation findings are used to advocate a particular standpoint. Our data have, nevertheless, shown that if a person or a group of people use an evaluation to support their aims, the evaluation is more likely to lead to impacts. This finding has been made even in studies on the use of research results in decision-making. Research and evaluation information tend to become "ammunition for the side that finds its conclusions congenial and supportive" (Weiss 1979).

This kind of *advocative* use of evaluation results is a very important precondition for impacts. Evaluations alone do not have sizeable impacts. Likewise, if an evaluation recommends changes that powerful groups disagree with, they are not likely to be implemented. The evaluation is simply ignored.

An example of such advocative use would be provided by the evaluation of experimental nuclear and high-energy physics in Finland. It included comments on a planned accelerator. The report was, however, very cautious, and it was consequently interpreted and used in opposite ways by both the advocates and opponents of the accelerator. The end result was a compromise, namely, an accelerator of a more moderate size than the one originally envisaged.

The impacts of evaluations tend to be rather small-scale. If a larger change results from an evaluation, an option for change must have received strong support before. The evaluation may have had a contributory effect and been used by advocates for specific purposes, but it is difficult to say whether the same end result would have been achieved if the evaluation had not been made.

There are other factors which affect the impacts of evaluations. Such factors include the *relevance* of the issues that the evaluation addresses to the concerns of decision makers. When evaluations were started in the Nordic countries, research councils were not well aware of the possibilities offered by evaluations. Thus, in

Norway and Denmark, evaluations of research fields started as an exercise "to gain experience in evaluation" (nuclear physics and nuclear chemistry in Norway and crystallography in Denmark). The fields evaluated were chosen on the basis of being "suitable", e.g. suitably large or international, not because there were special problems or needs in the fields in question. Later, after the research councils had gained more experience of evaluations they became increasingly aware of the importance of clearly defined evaluation aims. Still, in a large number of cases, evaluations have lacked such aims.

Even if evaluations had been started with clear aims in mind, this is not enough. The aims have to be *communicated* to the evaluators. There is often insufficient communication between those who commission the evaluation and the panel. We referred to the example of the planned accelerator in Finland. The experts arrived in Finland without having been told beforehand that they were expected to assess the plans. They did not obtain sufficient information about the plans, and consequently made only very guarded remarks on the subject. These were subsequently interpreted in various ways depending on the interpreter and his interests.

Interaction and communication between evaluators and those who are its potential users is important even after an evaluation. Special efforts in *dissemination* are needed to enhance the implementation of evaluation findings. Research council organizations still pay far too little attention to the follow-up of evaluations, be it a question of disseminating the findings to other decision-making bodies or to the scientists who have been evaluated. One example of this is the fact that in many cases evaluation reports have quite a restricted distribution. Scientists at large in the research field do not even get a copy even though they are expected to act upon the evaluation. It is in fact amazing that relatively large sums are invested in evaluation, but so little attention is paid to the utilization and follow-up of the results.

As far as evaluated scientists are concerned, it would often be very helpful, if a workshop were arranged where they can discuss and argue the findings. It is good for the general work climate and morale, but also for finding new solutions to the problems the evaluation has pointed out.

With the exception of the Academy of Finland, it is not a rule to give evaluated scientists an opportunity to correct potential factual errors in the evaluation reports. This causes unnecessary frustration and decreases the credibility of evaluations among the scientists and scholars who have been evaluated.

Follow-up evaluation has only been implemented in Sweden where the Natural Science Research Council evaluates all research fields on a routine basis, and is on a second round. Elsewhere, we have not been able to detect that systematic attention has been given to evaluation by a second round or follow-up evaluation.

Impacts on evaluated scientists and scholars

The impacts of evaluations on scientists and scholars are usually indirect and not easy to detect. They occur via measures taken by the research councils or other funding agencies, for example, through their decisions on funding, through the introduction of new modes of funding or new criteria to be applied in decision-making. These decisions and new policies affect scientists' behaviour indirectly while they strive to comply with them to attain research money.

Scientists' and scholars' observations of the effects of evaluations, for example on funding decisions, do not correlate with the results of evaluations; negative or positive evaluation does not automatically lead to negative or positive impacts in terms of research money, promotion, etc. There are, of course, exceptions. The observed low correlation between evaluation and impacts often leads to a conclusion that evaluations do not have effects at all.

Evaluations have reinforced tendencies which had started earlier. For example in the 80s, Denmark and Finland have experienced changes in scientists' publishing behaviour towards more publications in English and in refereed journals. There has also been an increase in competitiveness and scientists' international collaboration. Evaluations have played a role in these developments, but they cannot account for all the changes.

Only very few interviewees admitted that evaluation has helped them reorient their work; most claim that they know best what they should do, or alternatively, they were aware of potential problems in their research or of new avenues for future work. Evaluations of large entities, such as whole research fields and research institutions, rarely go into such detail that they could comment on the orientation of projects, or if they do so they often do not do it with sufficient expertise. A small panel of experts cannot cover all lines of research within a broad area. Consequently, the discussions the panel has with evaluated scientists rarely concern issues of substance (either in theory or method) in research work.

The arrangements made for the site visits also discourage open and in-depth discussions. There are often too many people present, for example, all research groups from a particular department plus departmental heads are gathered in the same room and witness each others' interviews. Besides, the discussion often takes the flavour of an interrogation which is not conducive to a fruitful dialogue between the panel of experts and the evaluated scientists and scholars.

Positive evaluation enhances scientists' status or gives them moral support. Very often this happens to those who have already been noted for their work, who have previous good connections (even with the evaluation panel), and who are able to communicate with the foreign experts, that is they have previous experience of participation in international scientific meetings and person-to-person contacts.

Evaluation, therefore, reinforces the status of those who already have been noted for their performance or who have a high status. It is thus conservative.

Nevertheless, there are examples in Finland where even junior scientists have been able to benefit from evaluations by getting a good review, encouragement and subsequently large research contracts. This is because they had been evaluated as independent scientists in spite of the fact that they do not have tenured positions. In other countries, especially in Denmark, such scientists are not likely to meet an evaluation panel. This is partly due to the fact that career structures differ; in Finland scientists without tenured positions can function as leaders in projects funded by the research councils while this is not usually the case in Denmark, especially in basic research. We have noted that evaluation is potentially a more beneficial experience for junior scientists. Meeting highly qualified and knowledgeable experts is more stimulating and encouraging for them than for senior scientists.

Negative effects

Evaluations have unwanted effects which are negative for individual scientists or research environments. For example, evaluations upset the work climate and cause uncertainty in research institutions until the evaluation report comes out and potential changes in scientists' positions or work environment are observable. This is a short-term effect which is not necessarily bad in the long run.

There are other, potentially more harmful effects over time. We have found that scientists expect criticism to help them reorient and plan their work for the future. However, a negative evaluation, in spite of opposite expectations, often decreases work motivation. This is so especially if it is very negative as was the case with some reports. It is not easy to receive strong criticism concerning the work in which you have invested a lot of your energy and other resources. If strong criticism is levelled at a university institute, it may spoil the work climate and decrease work motivation for a decade or more in that unit, unless new resources, such as personnel, equipment, etc, are provided to renew ideas and reorient work and work habits. It is easier to reorient work by employing new people than by forcing old ones to adopt new ideas. The research councils should carefully reflect how they could use evaluation findings in a constructive way, without detriment to research environments.

Negative evaluation also causes fears of a misuse or a negative use of evaluation, such as labelling and firing people, cutting off resources, giving undue attention to negative comments in evaluation reports, etc. As some examples prove, such fears are not unfounded.

Last but not least, there is a danger of an undue emphasis on one-sided criteria of performance in evaluation. This will lead to negative impacts on the overall health and performance of academic institutions. For example, we have noted a

strong emphasis in Denmark on the use of the number of articles in refereed journals as a major criterion in evaluations. This has been evident both in the evaluations of research fields by the research councils and the control mechanisms universities have introduced. Such an emphasis tends to undermine the teaching function of academic institutions; in addition the popularization of research results and provision of expertise are viewed negatively, since they take time out of activities that bring more merit. The effect of such an undue emphasis is not equally harmful in all fields; still, by and large, it is the most worrisome effect of evaluations in the long run.

References

Luukkonen T. & Ståhle B. Quality evaluations in the management of basic and applied research. *Research Policy* 19 (1990) 357-368.

Weiss C. The many meanings of research utilization. *Public Administration Review* 35 (1979) 426-431.

Odd Nakken

A Comment on the Follow-up and Use of Evaluations

Plans for the follow-up and use of evaluations should be an integral part of the evaluation itself. The reason for doing an evaluation is that we want to use it. But what do we want to use it for? The answers to this question should be made quite clear before evaluations are prepared and started. Nothing is more frustrating for people than to experience that evaluation reports are filed away or make no impact. Workers (scientists and others) should be confident that "something for the benefit of my own situation/for the benefit of my own organization" will be the outcome. Hence the time factor is essential: The lag from the time an evaluation report is completed until the follow-up is completed or at least commenced must be as short as possible.

The Institute of Marine Research (IMR) was evaluated in 1986/87. The main conclusions in the evaluation report delivered in October 1987 were:

- IMR should get a board of directors
- IMR should be reorganized
- IMR should increase its competence within certain research fields.

The report described both the new organization as well as the measures that should be taken in order to achieve increased competence.

During the evaluation process communication between the evaluators and the employees was extensive and a vast majority of the employees agreed with the report's conclusions. It was accepted and expectations were high among the staff at IMR.

Now, the reorganizing had to be handled by the new board of directors and it took two years before the government (the Ministry of Fisheries) appointed the members of the board. The board worked fast when it eventually "came to power", but during the two long years that passed without any action or reorganization at all, expectations decreased and frustration increased among IMR staff. However, while waiting for the government to appoint the board, IMR carried out quite a substantial increase in its competence; both the number of PhDs and the quality of research support activities were increased considerably.

The lesson I learned from this was: Following the conclusions and recommendations in the evaluation report, there should have been a time table showing when the different steps in the reorganization were to be completed as well as a statement of the necessity to keep to this time table in order not to lose support in the organization.

Morten Staude

A Comment on the Follow-up and Use of Evaluations

Introduction

A few brief observations concerning the follow-up and use of external independent evaluations - as seen from my point of view - implying

- that I do not mean to present *general* conclusions, that I do not mean that it always has to be like this,

- that I will be talking about the past, not about the future.

Evaluation: A fashionable theme. If we are not careful, it may well disintegrate into being nothing more than a slogan, or a ritual.

The system lacks reliable priority-setting mechanisms: A role for external independent evaluations? As seen from my perspective it is doubtful if external independent evaluations (ex post) have been or will ever be - a useful vehicle for this purpose.

Evaluation: Value-for-money? I will come back to this in a moment.

Usefulness: (Maybe) to science policy apparatchiks, to science policy spectators and to media - but not to the R&D community.

Follow-up/use

As I see it the follow-up of external independent evaluations has been *rather limited* - for various reasons. The responsibility - or the blame, if you like - has to be shared. There seems to be scope for improvement both on the supply-side - on behalf of the evaluators and on the demand-side - on behalf of apparatchiks, customers, decision-makers commissioning evaluations.

Supply-side issues:

- Norway is a small country, especially when talking about qualified R&D evaluators.

- The format of the evaluation report should be seen in a policy context - the more operational, instrumental, action-oriented the easier for science policy makers to take on board.

Demand-side issues:

- The basis for carrying out useful evaluations (ex post) is often rather insufficient, where policy makers ask for en expert evaluation ex post of something that - as a fact of life - was rather vague, even ex ante.

- There is also scope for improvement with respect to the evaluation mandate format. As I see it we would benefit from being more precise in formulating the mandate - not trying to cover everything, not trying to fool ourselves by acting as if the mandate and the evaluation budget can be decided separately.

- Even if everything else is fine, the ability and willingness - of those commissioning the evaluation - to implement necessary changes/adjustments in response to the evaluation is often rather limited. Fading enthusiasm on behalf of the decision-makers.

Finally two points that may be rather dubious:

- External evaluations can sometimes be seen as a first line of defense - initiate an evaluation to keep your paymasters happy (or quiet).

- In general we are not inclined to learn from other people.

To conclude: What has been said is not primarily a criticism of those carrying out evaluations (ex post), but rather a criticism of the apparatchiks asking for an evaluation of something that - as a fact of life - is rather difficult to evaluate properly, and who are also showing insufficient enthusiasm, both in the planning and in the follow-up phase. There is ample room for improvement.

Research councils should be well equipped to commission external independent evaluations (ex post), but ought to be more selective and also more serious about it - to achieve value-for-money. One also needs a discussion of external independent evaluations (ex post) vs. other means of quality assessment/quality control. Based on experience, as indicated above, I think research councils and other bodies commissioning R&D evaluations would benefit from redirecting their evaluation work from external independent ex post evaluations to real-time evaluation/monitoring using internal resources.

Hans Skoie

Research Evaluation - What Should the Research Councils Do?

Madam chairwoman, ladies and gentlemen. I am really in doubt if you and most of us can stomach any more now on a Friday afternoon. So, I will try to be brief. As many good things have been said, there is not much left for me to address.

First, what is my credo regarding evaluation? What is my basic belief? My assumption? I think evaluation has arrived and should be with us in moderate amounts. We need evaluation as a corrective and for all the good reasons you have given here during this conference. But it should not be launched on a grand scale. The resources involved for a good evaluation are sizeable and the doubt we may have of the accuracy and effect of an evaluation should not be discounted.

I also think it is better to do few evaluations "up to standard", i.e. professionally rather than many of dubious character. Unfortunately, a lot of what has been done in Scandinavia in this area so far has not been sufficiently professional. A consequence is lack of credibility, as Terttu Luukkonnen has pointed out, and accordingly, lack of impact. The evaluation of research may change the reputation of teams and individuals. If there is anything scientists do care about, it is professional reputation. That is important in research councils as well. Often it is not the way you write a proposal, but who has signed it, who sponsored it and what does the publication list look like? This is of particular importance for basic research.

Furthermore, I support a selective procedure. The pattern for oral exams in Norwegian gymnasiums may serve as an analogy. All pupils are entitled to take oral exams, but who does is decided by a lottery. You will have to be prepared for an oral exam which may or may not materialize. I think this keeps everybody on their toes in this system. Accordingly, researchers may be up for evaluation and should be prepared for that.

Assessing research quality and the results of both a scientific and non-scientific nature is the essence of any research evaluation. I take issue with all of those who just talk about relevance. In applied research and development actual results and the non-scientific goals stated in the research outlines are usually what should be looked for - not only "vague relevance".

The question of process is also of importance in certain cases and for certain purposes. Process evaluation of programs, for example, has been addressed earlier today. So, both should be done.

Turning to the research councils, why, should research councils get into research evaluation? They are major actors in research funding and research policy in most of the Western type of countries. How strong they are may vary somewhat from country to country. I think I dare say that in Norway the research councils probably have a stronger position than in most other Western countries. The research councils are national bodies, and we do not have any other similar national bodies. One might argue that we should establish something new as an alternative to research evaluation. The councils may often have something at stake. They often made the choice of supporting that project, program or institute, etc. To put it bluntly, they could be interested in "covering up" their mistakes, "hiding" that they were not up to making the best choices, at least in hindsight. For a small country, however, it is too expensive to build up an alternative. Accordingly, we have to stick with the research councils having an evaluation role. That role should also be a national role much broader than the councils have actually funded themselves.

There is, however, a second type of argument against the research councils. So far they have not a particularly good record in this area in the Nordic countries. I am not especially impressed with what they have accomplished until now.

Let me address some weaknesses as I see them. First of all they have often been weak on methods. The Swedish Natural Science Council's method for basic research is helpful as far as it goes. But there have been several "blunders" as pointed out by Per Seglen, for example. The medical faculty at the University of Oslo is not the only one.

Furthermore, the method applied may be good in other cases, but the efforts are not deep enough. The resources allocated for an evaluation may not allow for sufficient data collection, expert judgements, time, etc.

Concerning the use of evaluations, it is my opinion that many of the councils have also been weak on this point. This includes follow-up with regard to the researchers involved. Professor Rekstad is right in blaming the natural science research council in Norway for good reasons on this account. However, at the same time he seems to have an assumption that since the researchers came out of the evaluation with flying colours they should automatically get extra resources. Any man or woman used to making up budgets would object to such reasoning and may say they "now have a good reason for keeping up the high spending level on this item also next year". In Norway it seems to me that NORAS has done the best so far in its evaluation efforts. They also have weak points. But still, I see its research evaluations as probably the best so far among the Norwegian research councils. An experimental attitude toward this difficult area may explain that the Council seems to have got a better grasp of the complex problems involved.

What should be done? First of all, I would like to point out that research administration, the research councils, the program committees, etc., should get a

better grasp of research evaluation. They have to realize the difficulties and complexities involved. That has not always been the case. By the same token, what is needed is much stronger professionalization of the evaluation efforts than usually has been the case so far. To do that, some guidelines for different types of evaluations are desirable. The various evaluation tasks may require very different methods. The tendency to appoint an academic panel has been much too widespread, for example. The guidelines should also point to common pitfalls. I think such guidelines might be a great help in most types of evaluation work.

It is very important that evaluation methods are carefully discussed and adjusted according to what the evaluation is all about - the task. The other day we heard Lord Flowers, an experienced Englishman with a strong research council background, make harsh statements about research evaluation. However, it turned out that what he meant was research evaluation of a particular kind in basic research, which is only 10 to 20% of all R&D. At this seminar I have also noticed similar statements which do not specify sufficiently what kind of research we are dealing with. Is it basic research, applied research, development? Are we talking about programs, institutes, or research organizations, etc.? By the same token, it follows that the evaluators' tasks should be spelled out as clearly as possible in the mandate for the evaluation.

I liked what Mr. Massimo pointed out this morning, namely the need for "supplementary studies". The judgement of a panel may not always be sufficient. I have in mind supplementary material and studies which may be considered as appendices to panel reports. Here special studies or statements by experts could be included which only a single person or institution may be held responsible for, not the panel as such. Supplementary material of this sort may be very valuable and I am glad to learn that the European Community uses this practice. As often as possible such material should be presented separately and not under the general responsibility of the secretariat because the secretariat is under the control of the chairman and the panel. Of course it is the responsibility of the panel to say 'yes' or 'no' to including such studies. Do they accept them as decent studies? If so, they should be included. I will once again underline that I think we have gone too far with the panel approach in Norway. That seems to be the only way of doing evaluations so far. I also think that to have a single expert present a review paper on a particular field may sometimes be preferable. If you have to put your reputation on a paper like that, you are very careful about what you write and the effort you put into the article - you don't write it on the airplane from Oslo to Berlin. A committee may often compromise and often no one really feels responsible for the wording. To individualize responsibility like this may sometimes be of great help. And, this method may be much cheaper than panels. In practice, a combination may be particularly worthwhile.

The tendency to use consulting firms in this area has gone too far in this country. I have noted with great interest that the Ministry of Petroleum and Energy recently made an interesting remark on this point. In a statement to the Grøholt Commission, the Ministry criticized research councils for extensive use of consulting firms in evaluation work. "That we can do as well, and we can do it directly. We have assumed that we have research councils because they are experts on research - as the consulting firms are not. We don't need a research council which goes to consulting firms. That we can do ourselves." That reaction is quite interesting. The Government of Norway established the research councils to be experts on research, and not taking on the job, but just leaving it to amateurs who are in banking one day and somewhere else the next day, is a doubtful practice. Of course consulting firms may be used occasionally, particularly on the management side. I can see that, but they have been used too broadly in Norway. That is my view.

The councils should try to obtain an evaluation effort of a cumulative nature. Unfortunately there is the opposite tendency, an ad hoc approach. Each evaluation seems to start from scratch. I'm sure Sweden does not do it like that in the natural science research council, because it has guidelines and a worked-out policy in this area. But in many other cases I have seen a tendency to an ad hoc approach all over Scandinavia. You even find examples of people being appointed to panels who did not behave professionally the last time they were assigned a similar task. The cumulative aspect is not sufficient in my view to really build up expertise with regard to evaluation. The direction of the National Science Foundation seems the way to go. They have a unit particularly dedicated to this kind of work in the Director's Office.

The councils should explicitly aim at always establishing a fair evaluation process. That is important, also in order to gain credibility among researchers and policy makers as well. Scientists should be given ample opportunity to give evidence during the evaluation process.

An opportunity to correct factual errors in the evaluation report should always be given, and I am sorry to hear from Terttu Luukkonen that this is not always the case in the Nordic countries. Much misunderstanding could be avoided by such an effort.

I do think the reports should be discussed openly before councils or other bodies act on them. Arranging hearings or seminars may be appropriate. The Swedish evaluation of sociology some years ago led to much discussion of that kind. Actually I was quite impressed with that evaluation which also included several separate studies as appendices and illuminated the field of sociology in Sweden to a great extent.

The evaluation report should be made public, if the council finds the report to be of a decent professional standard. And I can assure you that I know of more than one report in Norway which probably should not have been published on this account. A research agency should not publish rubbish of this kind; that would be counterproductive and unfair. However, the minimum should be professional reports which can be published.

Finally a remark on criteria. Terttu Luukkonen mentioned the tendency to one-sidedness in some of the evaluations she has looked at. That often seems to be the case and should also be addressed in the guidelines. We actually did it in Norway when we presented guidelines for the social sciences some years ago. First of all there often seems to be an academic bias. Panels are often overstaffed with university researchers even in cases where applied work is the major task for evaluation and such staff tend to use academic criteria too heavily. Furthermore, the entire effort under evaluation and types of publications or otherwise should be listed for the research groups under evaluation, i.e., classical academic papers as well as other types of reports and presentations - and the assessment made explicit according to various criteria.

Concerning universities which are strongly influenced by disciplines, you should be careful to address the entire activity in the evaluation, i.e. both the educational side and the research side. And that may now be done in Norway where the Ministry of Education, Research and Church Affairs has also taken an interest in the evaluation of education. A simultaneous approach like that means that people can't say: "I am so good at teaching" and the others say "I am so good at research". You should aim at getting the total picture of a department. This approach may also be appropriate for applied institutes outside academia. They have other types of work than research. It may be appropriate to ask how well they do the total job including the research-related work which is also part of the professional work at that institute.

My last point is to repeat that the research council system should have a major role in research evaluation. Within a rather moderate evaluation activity, this should include responsibility for improving evaluation methods and what you might call real-time evaluation of evaluation activities, as for example, Per Seglen pointed out.

Egil Kallerud

A Comment on Research Evaluation - What Should the Research Councils Do?

Research evaluation has of course been a core activity within research councils since their inception. However, the 80s have seen a surge of new forms of research evaluation, requiring innovation in both organization and methodology. Seen as an expression as well as an agent of general trends in research policy development during the 80s, the evaluation "fad" may have challenged some of the traditional ways of doing research policy within research councils, in terms of accountability of results, transparency of processes, justification of decisions and outside participation in deliberations. The complexity of research evaluation by and within research councils has increased as more variables and more actors become part of the game.

This is clearly born out when reviewing experience gained so far with so-called *ad hoc* evaluations, i.e. specific and major evaluation efforts of programmes, scientific fields or institutions. Independent, *ex post* evaluations have so far not been used systematically for purely accountability purposes, in contrast to Sweden (Gidefeldt, this volume), nor are performance data used as integral parts of the management of research within research councils. Skoie focusses on the preconditions for more professionalism in research councils in setting up, performing and using such *ad hoc* evaluations. In fact, this lack of professionalism may be a sign that a transition is taking place, challenging and extending the traditional roles and functions of research councils.

Which has, of course, basically been that of *gatekeepers*, i.e. to regulate access to the system and its scarce resources by peer review-based decision processes. Thus, "evaluation" has mainly come to mean "appraisal" of grant applications. As evaluation is also expected to account for the effectiveness, productivity, and quality of research *post hoc*, or even, to an increasing extent, the impact of research on society, the immaturity and underdevelopment of evaluation methodology become evident. *Ad hoc* evaluations, available as separate reports that describe activities, assess resources and results, and propose actions, are - in terms of openness to scrutiny and criticism – clearly very different from mail reviews that are usually withheld from public scrutiny, and the processes engendered by such evaluations are different by nature from those that take place behind the closed doors of committee, council and board meetings. The professional management of these kinds of documents and processes is one of the important challenges raised by *ad hoc* evaluations. Skoie is right when he emphasizes the importance of ensuring the

fairness of the process, the factual correctness of information, the public availability of the report, and the participation from parties directly and indirectly involved. Blunders committed by my own council emphasize these points.

The higher stakes of major evaluation efforts, compared with the incremental changes in resource allocations that usually result from research council decisions, is another aspect indicating that the task at hand is different by nature, not only by degree. Conflicts of interests will often be intense, and decision making difficult. Depending on the extent of the transformations that one hopes to achieve by the use of *ad hoc* evaluations, there is a possibility that they will expose the ambiguity of the role of research councils and their limitations as research policy bodies.

Norwegian research councils, or at least most of them, are not yet beyond the experimental or explorative phase in their use of *ad hoc* evaluations. It is about time that they sum up experiences made so far and try to specify some basic guidelines for future evaluations. I agree with Skoie that guidelines for *ad hoc* evaluations should be worked out. The one "handbook" recently produced by the council for agricultural research, raises one question: to what extent should these "handbooks" be overviews or "general introductions to research evaluations", listing and commenting upon possible options and situations for all types of research evaluations? I think they should mainly be seen as part of the process towards the *standardization* of each council's use of *ad hoc* evaluations, both specifying criteria for the selection of evaluation types seen as appropriate for that council, and laying down precise rules for organizing the types of evaluations selected. The applicability of such guidelines could be enhanced if explicit comments on earlier experiences are included.

The professionalization of research council performance of evaluations may be facilitated if extensive support is given by the Institute for Studies in Research and Higher Education. The Institute is expected to perform (parts of) evaluations itself, and act as an advisor to the councils. The Norwegian Research Council for Science and the Humanities, NAVF, should and will do what it can to enable the Institute to improve its competence and enlarge its capacity for those tasks.

Ad hoc evaluations are, however, not the only issue that should be addressed when discussing evaluation by and within research councils. Another is the relationship between the use of *ad hoc* evaluation and each council's general routines for the monitoring and internal evaluation of supported research. As the name indicates, *ad hoc* evaluations are not supposed to be the normal procedure for the evaluation of all research. As *selective* efforts, they should be seen as particular measures within a specific context, having a clearly stated purpose and expected function within a particular decision process. Of course, general accountability could be the main purpose of independent or external evaluations. I do not think, however, that the Swedish system described by Gidefeldt is likely to be adopted in

Norway, at least not without stipulating some (rather high) threshold for resources involved, below which the costly procedure of external evaluations should not be applied.

Ad hoc evaluations can be seen as *complementary* to monitoring and internal evaluation. More systematic recording and analysis of information on supported research projects might reduce the need for *ad hoc* evaluations, and make them cheaper when performed. Better reports, use of interviews and site visits could provide what is needed in most cases. Incidentally, some blend of internal and independent evaluation has recently been introduced in NAVF, intending to use international peers to monitor progress and evaluate results of projects/groups that receive large, annual block grants for several years.

All research councils have the responsibility to improve their monitoring and internal evaluation systems. This, however, may mean different things, depending, e.g., on whether the activities monitored are those only supported by the council or a national activity within whole fields of research. What makes Norwegian field evaluations different from the Swedish system described by Gidefeldt, is actually the national scope of our evaluations. This distinction relates to the extensively discussed issue in Norway of the national "strategic and evaluating" functions of research councils. Research councils are expected to play a crucial part within the Norwegian research policy system, perhaps more so than in most other countries. No doubt, the research councils could and should play a substantial national,"strategic" role, and broad responsibility for research evaluation is certainly an important part of that function. However, it is necessary that appropriate conditions are established for being efficient agents of "strategic functions".

One problem is that it is possible for NAVF to perform field evaluations and use information on national output and productivity in whole fields as part of the council's own planning process. Actually, that is one important way to enhance the function of research evaluations within the council. The impact of such studies will, however, be limited to the small proportion of the total research expenditure that is actually controlled by the council itself. Nevertheless, the *strategic* functions of the research councils could and should be extended beyond that.

The problems that each council has to face in trying to define an extended strategic function will vary from council to council; one reason for this is that they have to negotiate with different ministries. In some of the applied research councils, this may possibly not be a problem, e.g., when the Norwegian Research Council for Applied Social Science, NORAS, is asked formally by a ministry to evaluate institutes owned and run by that ministry. The problem may surface, however, when some research councils take a stand on general research policy issues, although not expressly asked to do so by the authorities responsible for the activities in question. Examples are the discussion on the reorganization of institutes for technical research

initiated by the Royal Norwegian Council for Scientific and Industrial Research, NTNF; and possibly the corresponding review by NORAS of the organization of social science institutes. (My point is, of course, not invalidated by the fact that neither of these reviews are (based upon) *evaluations* in the ordinary sense. Their prospects of success might probably have been enhanced if they had been). A similar problem for NAVF is the consequence of what seems to be little interest for university *research* within that section of the Ministry for Education, Research and Church Affairs which is responsible for allocating general university funds. One might question the assumption that the system of "result-oriented planning" introduced in higher education institutions will be able to handle the complex problems involved in the overdue task of enhancing accountability for this part of public research expenditure. It is certainly not possible for Norway to copy either the evaluation system adopted by the British Universities Funding Council, UFC, the Dutch "conditional financing" system or the French "contracting" system. One general lesson to be learned from these examples might be, however, that the Ministry responsible for university research has to assume an active and orchestrating role. The Ministry could assign a clearly defined, advisory role to NAVF, and be prepared to act on its advice. Of course, in the absence of an active Ministry, the institutions themselves might ask the Council to organize independent evaluations of faculty research. The point is that research council evaluations as a rule should not address activities and propose measures outside the scope of the councils' own authority, except in agreement with the agencies responsible for that activity. If this requirement is not fulfilled, broadly oriented evaluations will probably be shots in the dark.

One should, therefore, be careful when trying to implement Skoie's recommendation that the evaluation of research and teaching should take place simultaneously. Research councils should not undertake the evaluation of teaching activities, except when there is an agreement with the proper institutional or national authorities. I do not, for example, feel quite sure that the NAVF subcouncil for the humanities has clarified how its planned evaluation of both research and teaching in university departments of English is to be followed up.

Finally, I want to emphasize that research councils should see evaluations not only as inputs to decision making, but also as opportunities to gain better knowledge of the conditions, functioning and impact of current research. That knowledge may have indirect usefulness as important as that which stems from relevance to particular decisions. NAVFs Institute for Studies in Research and Higher Education should be allowed to – and want to – exploit material collected and experience gained from evaluations to contribute to general science studies, and make accessible the insight acquired from these studies to research council staff.

List of participants

Antonsen, Inger, Research Coordinator, The Agricultural Research Council of Norway.

Apeland, Jacob, Professor, Department of Dairy and Food Industries, Agricultural University of Norway.

Berge, Arne, Research Coordinator, The Joint Board of the Norwegian Research Councils.

Brofoss, Karl Erik, Research Coordinator, Institute for Studies in Research and Higher Education, The Norwegian Research Council for Science and the Humanities.

Buflod, Halvdan, Research Coordinator, The Norwegian Research Council for Science and the Humanities.

Daling, Liv, Head of Division, The Norwegian Council for Fishery Research.

Denstad, Brit, Director General, The Norwegian Research Council for Applied Social Science.

Edvardsen, Erik, Controller, Royal Norwegian Council for Scientific and Industrial Research.

Ensby, Simen, Director, Royal Norwegian Council for Scientific and Industrial Research.

Foss Hansen, Hanne, Associate Professor, Center for Public Organization and Management, Copenhagen Business School.

Gidefeldt, Lars, Head of Project Department, Swedish Natural Science Research Council.

Guy, Ken, Senior Fellow, Science Policy Research Unit, University of Sussex.

Hallén, Arvid, Director, The Norwegian Institute for Urban and Regional Research.

Hammerqvist, Sten-Erik, Research Coordinator, The Norwegian Research Council for Applied Social Science.

Hauknes, Johan, Coordinator, The Joint Board of the Norwegian Research Councils.

Helgesen, Gro, Deputy Director, The Norwegian Research Council for Science and the Humanities.

Hilmen, Anne-Lise, Acting Director General, The Norwegian Research Council for Science and the Humanities.

Irgens-Jensen, Synnøve, Research Coordinator, The Norwegian Research Council for Science and the Humanities.

Isaksen, Randi, Secretary, The Norwegian Research Council for Science and the Humanities.

Jordell, Karl Øyvind, Research Coordinator, Institute for Studies in Research and Higher Education, The Norwegian Research Council for Science and the Humanities.

Kallerud, Egil, Research Coordinator, The Norwegian Research Council for Science and the Humanities.

Kyvik, Svein, Research Coordinator, Institute for Studies in Research and Higher Education, The Norwegian Research Council for Science and the Humanities.

Larsen, Ingvild Marheim, Researcher, Institute for Studies in Research and Higher Education, The Norwegian Research Council for Science and the Humanities.

Lindgren, Bjørn, Research Coordinator, The Norwegian Research Council for Applied Social Science.

Luukkonen, Terttu, Researcher, The Academy of Finland.

Løvland, Jarle, Director of Research, Norwegian Institute of Fisheries and Aquaculture.

Massimo, Luigi, Head of Research Evaluation, D-G XII, Commission of the European Communities.

McCullough, James, Director, Program Evaluation Staff, National Science Foundation.

Nakken, Odd, Director General, Institute of Marine Research, Norway.

Nereng, Berit, Research Coordinator, The Agricultural Research Council of Norway.

Nybom, Torsten, Lecturer, National Swedish Board of Universities and Colleges.

Olsen, Tore, Director General, Department of Research, Ministry of Education, Research and Church Affairs, Norway.

Olstad, Finn, Research Coordinator, The Norwegian Research Council for Science and the Humanities.

Rekstad, John, Professor, Department of Physics, University of Oslo.

Roll-Hansen, Nils, Professor, Institute for Studies in Research and Higher Education, The Norwegian Research Council for Science and the Humanities.

Seglen, Per, Professor, Institute for Cancer Research, The Norwegian Radium Hospital.

Sivertsen, Gunnar, Researcher, Institute for Studies in Research and Higher Education, The Norwegian Research Council for Science and the Humanities.

Skavlan, Bjørn, Research Coordinator, The Norwegian Research Council for Science and the Humanities.

Skoie, Hans, Head of Department, Institute for Studies in Research and Higher Education, The Norwegian Research Council for Science and the Humanities.

Smith, Keith, Research Director, Innovation Studies and Technology Policy Group, Norwegian Computing Centre.

Staude, Morten, Director, Royal Norwegian Council for Scientific and Industrial Research.

Søgnen, Randi, Researcher, Institute for Studies in Research and Higher Education, The Norwegian Research Council for Science and the Humanities.

Thornquist, Morten, Deputy Director General, The Norwegian Research Council for Applied Social Science.

Tærum, Eli Ragna, Research Coordinator, The Agricultural Research Council of Norway.

Tønder, Johan-Kristian, Director, Institute for Studies in Research and Higher Education, The Norwegian Research Council for Science and the Humanities.

Van Raan, Anthony, Professor, Centre for Science and Technology Studies, University of Leiden.

Vislie, Tone, Director, The Norwegian Research Council for Fishery Research.

Voje, Kirsten, Section Manager, Royal Norwegian Council for Scientific and Industrial Research.

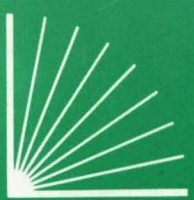
Walters, Sue Ellen, Language Consultant, Institute for Studies in Research and Higher Education, The Norwegian Research Council for Science and the Humanities.

Øberg, Stein, Research Coordinator, The Agricultural Research Council of Norway.

Research Evaluation

This report contains the proceedings of a conference on research evaluation held at Holmenkollen Park Hotel Rica, Oslo, 30–31 May 1991. The aim of the conference was to discuss how to perform research evaluations of high quality. The target group consisted mainly of staff in the research councils responsible for evaluative work.

The conference was arranged by the Joint Board of the Norwegian Research Councils and the Institute for Studies in Research and Higher Education, the Norwegian Research Council for Science and the Humanities.



NAVFs utredningsinstitutt
Norges allmennvitenskapelige forskningsråd
Munthes gate 29, 0260 Oslo
Telefon (02) 55 67 00

Institute for Studies in Research and Higher Education
The Norwegian Research Council for Science and the Humanities
Munthes gate 29, 0260 Oslo, Norway