

THE EVALUATION OF SCIENTIFIC RESEARCH: SELECTED EXPERIENCES

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Paris

59442

Document complet disponible sur OLIS dans son format d'origine

Complete document available on OLIS in its original format

FOREWORD

This document presents the proceedings of an OECD *Workshop on the Evaluation of Basic Research* which was held in Paris on 21 April 1997. The participants at the workshop included Delegates and experts from OECD Member countries, and the papers contained herein reflect the experts' views (and not necessarily those of their Member countries). This document contains a Summary and Conclusions and two structural parts which comprise the expert contributions: I) Country Overviews, and II) Institutional Experiences.

This workshop and its proceedings are part of the work programme of the *Group on the Science System* of the OECD Committee for Scientific and Technological Policy (CSTP). The Committee agreed to declassify this document at its 69th Session on 7-8 October 1997.

Copyright OECD, 1997

Applications for permission to reproduce or translate all or part of this material should be made to: Head of Publication Service, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France.

TABLE OF CONTENTS

SUMMARY AND CONCLUSIONS	5
PART I: COUNTRY OVERVIEWS	11
CHAPTER 1. EVALUATION OF SCIENTIFIC RESEARCH IN FINLAND	12
CHAPTER 2. EVALUATION OF SCIENTIFIC RESEARCH IN THE NETHERLANDS.....	27
CHAPTER 3. EVALUATION OF SCIENTIFIC RESEARCH IN THE UNITED KINGDOM.....	47
PART II: INSTITUTIONAL EXPERIENCES	59
CHAPTER 4. BIBLIOMETRIC ASSESSMENT OF RESEARCH PERFORMANCE IN FLANDERS	60
CHAPTER 5. RESEARCH EVALUATION AT THE CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (CNRS) IN FRANCE	74
CHAPTER 6. EVALUATION OF THE BLUE LIST INSTITUTES BY THE SCIENCE COUNCIL IN GERMANY	83
CHAPTER 7. RESEARCH EVALUATION AND UNIVERSITIES IN JAPAN: AN EXPERIENCE FROM THE UNIVERSITY OF TSUKUBA.....	91
CHAPTER 8. INTERNATIONAL EVALUATIONS OF THE SWEDISH NATURAL SCIENCE RESEARCH COUNCIL (NFR)	101
CHAPTER 9. UNITED STATES: THE EXPERIENCE OF THE NSF'S EDUCATION AND HUMAN RESOURCES DIRECTORATE	107

SUMMARY AND CONCLUSIONS *

Introduction

Research evaluation has emerged as a “rapid growth industry”. In most OECD Member countries, there is an increasing emphasis on accountability, as well as on the effectiveness and efficiency of government-supported research, as there is for many other categories of government expenditures. Governments need such evaluations for different purposes: optimising their research allocations at a time of budget stringencies; re-orienting their research support; rationalising or downsizing research organisations; augmenting research productivity, etc. To this end, governments have developed or stimulated research evaluation activities in an attempt to get “more value for the money” they spend on research support.

Despite the increasing prevalence of such evaluation efforts, the effectiveness of the various approaches to the evaluation of research has not been critically assessed. Nor has the question of the effectiveness of research evaluation been brought adequately to the attention of policy makers in most governments. On the other hand, the current state of the art in evaluation of research is based on specific methods and procedures that have been considerably enriched and refined in recent years and which deserve to be re-examined following the latest stocktaking made ten years ago by the OECD (see *Evaluation of Research*, OECD, 1987).

The 21 April 1997 Workshop on Evaluation of Research in Universities and Government-Supported Organisations — organised under the auspices of the OECD’s Group on the Science System (GSS) — was based on a series of nine presentations of concrete experiences, illustrating various approaches to research evaluation at both the country and the institutional levels. Although initially focused on the evaluation of basic research, the presentations, due to the very nature of issues at stake, covered a broader scope beyond “basic research” in the strict sense which, in fact, appears to be increasingly difficult to separate from applied or so-called “strategic” research. The contributions made at the workshop dealt more generally with the evaluation of research in both universities and government-supported organisations, such research being itself considered in relation to other functions performed in those institutions, e.g. training, technology transfer, etc.. A number of conclusions emerged from the workshop, which was attended by 70 experts and Delegates from OECD Member countries.

Evaluating the various levels of research systems

Research evaluation efforts can be focused on different entities differentiated by increasing levels of size and complexity. At least five types of such research entities have been the subject of evaluation, according to those experiences presented by the experts at the workshop. Sometimes entities at more than one level are evaluated at the same time. At the first level, evaluation can focus on the work of individual

* The conclusions prepared by the GSS Chair (W. Blanpied, United States) and the OECD Secretariat (J.E. Aubert), which are presented in this overview, are based on the summary made by the workshop rapporteur (M. Brennan, Australia).

researchers. Second, it can concern larger research groups, laboratories, and institutions such as universities. Third, evaluation can focus on an entire scientific discipline. Fourth, it can concern government programmes and funding agencies. Finally, evaluation methodologies can be applied to a country's entire research base. However, whatever the complexity and character of the entity being evaluated, it remains the case that research evaluation begins with the work of identifiable individual researchers. Thus, such individuals are bound to be either positively or negatively affected by the results of evaluations, including qualitative statements made in the course of such exercises. On the other hand, questions asked by different types of evaluation exercises usually vary according to the character of the entity being evaluated, even though such questions are (or at least should be) directed towards some form of quality assessment and framed with a view toward improving the management of the entities concerned.

The nine experiences presented by experts at the workshop can be summarised as follows, according to the different categories of entities involved.

The evaluation of researchers as individuals is illustrated by the French "Centre National de la Recherche Scientifique" (CNRS) experience where evaluation is performed as a basic instrument for personnel management and promotion. The Swedish Natural Science Research Council's (NFR) experience of international evaluation concerns both researchers and disciplines, while offering by the same token views on government support to specific people and projects. The Japanese experience shows how evaluation practices are being developed in universities as stimulated by recent government guidelines which aim to raise the level and significance of basic research conducted in the country's universities.

The German experience presents the methods used for rationalising the network of some 80 government R&D institutes, known as the "Blue List" institutes. The US National Science Foundation's (NSF) experience, dealing with evaluation of both university research and education performance, shows how that agency is attempting to comply with regulations requiring evidence of programme efficiency in all agencies of the US Government, including those involved primarily with the conduct and support of research. The Belgian Flemish Community's experience illustrates the evaluation of research activities in universities and government-supported laboratories in using bibliometric indications and in relating research evaluation to evolving funding mechanisms and broader priority-setting issues.

Finally, three overviews of evaluation practices covering an entire research system at the country level were presented by the experiences of Finland, the Netherlands and the United Kingdom. Although all three experiences aim at overall country assessments, they differ in the focus of the evaluation exercises themselves, as well as in the methods used and the nature of their feedback on policy making. The Finnish approach tends to focus on major instruments and institutions of the science system and has a long-term, although not necessarily direct, impact on policy choices. The British approach, of a more quantitative inspiration, is also more directly linked to the policy-making process. In the Netherlands, the aim is principally to stimulate evaluation efforts by actors themselves (universities, government laboratories, etc.) throughout the different parts of the science system.

Who evaluates? Internal and external evaluations

As suggested by the above summary, research evaluation can be performed by several different actors, depending on the objectives and specific context of the evaluation. That having been said, it is also the case that basically there are two major categories of evaluation performers. First, research evaluation can be a self-directed process when implemented by institutions themselves, whether they be major research organisations, universities, or funding agencies. Such efforts may respond either to self-discipline principles or to imposed government regulations. Second, evaluation can be performed by an external

organisation, either in response to specific government instructions or in compliance with more general rules. In the latter case, evaluation institutions have been established explicitly for such purposes on a more or less permanent basis. This case is most clearly illustrated by the contribution from the Netherlands, where several external institutions have received such a mandate.

An intermediate situation between internal and external evaluation occurs when institutions conducting their own (self) evaluation call on outside experts to perform it. A typical case is the use of foreigners as illustrated by the Swedish Natural Science Research Council which involves, systematically, the participation of at least three foreign scientists of international reputation (and only those specifically appointed), with support provided by the Council's own staff (plus an unbiased representative of the Swedish academic system) for implementing the work of those experts and producing related reports.

The priority objectives of the entity which calls for and conducts an evaluation, whether it be more an internal, self-evaluation process or an external, imposed one, appears to be less important than the pervasiveness of the evaluation practices throughout an entire research system, the soundness of those practices, and their impact on the decision-making process. From this perspective, there are important differences among OECD Member countries, as illustrated by the different experiences presented at the workshop.

What is evaluated? Outputs and outcomes

Of course, the primary results of research activities are the advancement of knowledge, and these results are usually referred to as the "outputs" of research activity. These outputs may take the form of publications, articles in scientific journals, books, conference papers, etc. Directly complementary to such scientific productions are so-called "second rank outputs" such as patents and related items that concern (potential) applications of research results. One may also include outputs such as designs, software development, etc., depending on the discipline involved.

Beyond such quantifiable "outputs", there are what are usually referred to as "outcome" indicators of research. These may include, for example: the production of graduates of high quality (including bachelors degree recipients, postgraduates, and research-trained graduates); concrete applications of research results, e.g. in the form of technological innovations; increased expertise and capacity of researchers and institutions for consulting; contract research services. Other types of "outcomes" may include international links developed by the research community under consideration (as measured by access to results, enhanced influence, etc.); and last but not least, the general contributions of research to culture. It appears that at all levels of evaluation, from individuals, to institutions, to entire systems, there is an increasing concern for including in evaluation exercises appraisals of outcomes, even though such outcomes are far more difficult, or even impossible, to define in purely quantitative terms.

How to evaluate: quantitative evidence as a basis for qualitative judgements

Whatever the subjects or level of the entities evaluated, and whatever the evaluation "culture" of the concerned country, research evaluation depends on two basic, complementary approaches: the use of quantitative indicators, such as bibliometrics, on the one hand, and the use of more qualitative peer judgements on the other. Only informed peers can express a judgement about the quality of fundamental research. Indeed, the notion of quality is so complex it cannot be grasped by quantitative methods which can only make some aspect of this concept visible. Generally, both are necessary, and tend to be used jointly. If there are clear caveats formulated in evaluation experiences for limiting the weight placed on

quantitative indicators such as bibliometrics, there is also broad acceptance that such indicators provide a measure of the “research production reality” that is usually irreplaceable. To a certain extent, it can be stated that “reality”, whatever it may be, does not exist if it cannot be measured in quantitative terms, and ought, at the very least, to be used to provide the quantitative background required to inform more qualitative peer judgements. Some research cultures are more concerned with the need for detailed output measures than others and pay particular attention to quantifiable indicators (see the UK contribution and the use of bibliometrics). In contrast, other cultures tend to limit the use of quantitative indicators (see the French and German contributions).

Areas such as research training which have so far been poorly documented in quantitative terms require special efforts. It would be highly useful to develop sound methodologies for quantification and tracing, e.g. of those holders of diplomas who have received research training and have left the academic research system. Such training indicators would be useful for managing, orienting and funding a country’s science base, to the extent that the training of new generations of scientists is seen as an outcome as important for evaluating a university as its research outputs *per se*, particularly from the perspective of industry.

The need for quantitative indicators — simple and communicative — is emphasised by the US contribution, to the extent that such indicators are indispensable in advancing the policy debate around clear issues. This is the sense of the activity launched throughout all US government agencies by the Government Performance and Results Act (GPRA) which was enacted in 1993 and which requires all government agencies to begin to assess their outputs beginning in 1998. An essential issue for agencies such as the NSF whose mission is to support basic research in universities and support science and engineering education at all levels is how to evaluate the outcomes of federal support for research and teaching. Such evaluations cannot be based entirely on quantitative output indicators. However, the various actors involved cannot avoid paying some attention to appropriate quantitative measures. At the same time, there is a broad consensus about the need to understand the relative weight to be assigned to such quantitative metrics. Above all, there should be no “fetishism” attached to different types of quantitative indicators.

Caveats

In a context of increasing concern for proper utilisation of government resources and expenditures, there is a need to be selective in both the number of evaluation exercises conducted and in the depth to which they are conducted. Workshop participants noted a phenomenon referred to as “evaluation fatigue” in certain countries in institutions subject to frequent, periodic evaluation. It was also emphasized that the financial resources required for effective evaluation need to be well understood from the outset, since it is often quite costly to launch in-depth measures of outputs or outcomes, which may, in fact, be only of moderate relevance for policy making.

A second set of caveats that emerged from the workshop discussion has to do with the impacts of evaluations on researchers and institutions. Over-evaluation unavoidably generates some measure of anxieties among the individual researchers concerned. There is a need for positive feedback from any evaluation exercise. In addition, the rules of the game should be made as clear as possible from the outset, that is: the criteria against which people, organisations, and disciplines are being judged; the processes by which they are to be examined; and the conditions under which the results are to be made public, and to whom. From this perspective, there are differences between practices in different countries, and it was suggested that the most transparent processes should serve as models.

Finally there can be a series of negative effects derived from evaluations performed in a manner which focus too much attention on research productivity, *per se*. For example, an overemphasis on research productivity in university-based evaluations can result in the neglect of the equally important training functions of these institutions. Likewise, evidence presented by some of the workshop experts suggests that overemphasis on research can lead to a temptation to exclude less productive developing countries' scientists from research facilities in the OECD countries. Both such results reduce the important roles of universities and government-supported research institutions in training new generations of scientists on the one hand, and in assisting in building up scientific capabilities in less well endowed countries on the other.

Finally, there are special problems involved in evaluating multidisciplinary work and emerging disciplines, which are often inappropriately appraised by current bibliometric countings, as well as by more qualitative measures based on peer judgements. Traditional research evaluation methodologies have tended to neglect the important so-called "grey" technical literature (the so-called "mode 2" knowledge), in favour of the codified production of knowledge in the form of publications in peer-reviewed journals and similar research outputs. However, such "grey" literature is often of cardinal importance for interdisciplinary work as well as for innovative developments. Finally, it was noted that most evaluation work is currently limited to hard sciences, but that the humanity and social sciences should not be neglected and should be subject to evaluation exercises as well, although methods and criteria need to be adapted.

Principal conclusions

Research evaluation is important and is increasingly viewed as essential in many OECD countries. But evaluation should not be considered as an end in itself. Rather, it should be developed and used more as a pointer to key policy issues and essential questions that need to be addressed. Research evaluation becomes useful to the extent that it helps in clarifying policy debates and moves decision-making processes forward on more rational and quantifiable grounds that improve the understanding of all partners involved in such decision making. In other words, evaluation should be conceived of and used primarily as a policy-making tool for managing different levels of research systems, rather than as a strict instrument of assessment and judgement, whether positive or negative. The evaluations should provide the basis for better decision making, by highlighting problems and formulating recommendations. But the evaluators should limit themselves to this objective and not be invited to try to replace the decision-making process which may take other considerations into account.

There is a need for multiple approaches to evaluation. In a colourful and appropriate metaphor offered by the UK expert, evaluation should be designed and regarded in terms of "tomography" with views from several different perspectives used to generate a complete, three-dimensional image. Each approach or method has both its positive features and its deficiencies. Each has been developed and refined to address a specific issue or type of assessment. But the limits of each approach should be clearly recognised. There is a need to search for, and implement, complementary assessment methodologies. With this in mind, the hazards of "religious wars" that have sometimes been waged between adherents of competing methodologies should be avoided at all costs.

Perhaps the most important conclusion to emerge from the workshop is that the evaluation of research at the institutional level must be conducted with full cognisance of the impacts of research on other, interrelated functions of that institution. For universities, such functions include: teaching and training, knowledge transfer to other social and economic sectors, international connectivity, and impacts on the broad national – and international – culture. More generally, effective evaluation of research outcomes at

the institutional level should be more concerned with the impacts of research activity on all functions of the institution, rather than on the evaluation of research productivity in its own right.

Evaluate evaluation!

A good deal of “meta-work” appears to be necessary in a context in which research evaluation exercises are likely to become increasingly prevalent in OECD Member countries. This work, which could usefully be undertaken by the OECD in line with its experience in science policy reviews, would serve several purposes: pinpointing advantages and drawbacks of evaluation efforts that are strongly influenced by national cultures; facilitating inter-country transfers of methodologies; and helping to focus evaluation efforts on most essential issues, while avoiding excessive financial costs.

Above all, the OECD may have an important role to play in raising the consciousness of the complexity and hazards of evaluation at the political and policy levels in its Member countries.

PART I: COUNTRY OVERVIEWS

CHAPTER 1. EVALUATION OF SCIENTIFIC RESEARCH IN FINLAND

Erkki Kaukonen, University of Tampere, Finland

Introduction: universities in transition

The increasing interaction between science and society accompanied by financial constraints in research funding have placed unprecedented demands on universities to prove their legitimacy and contribution to socio-economic development. To a large extent, these demands are mediated and formulated through science and technology policy guidelines and priority setting. The pressures towards accountability and efficiency are reflected in the recent boom of university evaluation, including research activities, which has emphasized the need to develop new kinds of institutional and external evaluations relative to the traditional ways of self-evaluation by the scientific community. The conceptual basis, adequacy and compatibility of different evaluation practices and criteria is a most complex issue arising in this regard.

Simultaneously, the role and status of universities as research sites is undergoing major transitory changes (Gibbons *et al.*, 1994; Kaukonen, 1997a). These changes are related to several factors and developments, which may be summarised as follows.

- ◇ Universities constitute just one institutional sphere of the national research systems. The other spheres, consisting of governmental (sectoral) and industrial research and development (R&D) activities, have generally outgrown the universities and in many countries – including Finland – carry out more than 80 per cent of total R&D activities.
- ◇ The restructuring of research systems has implied a clear shift of emphasis from the disciplinary context of knowledge production to an application-oriented and industrial context of R&D. In Finland – as may be typical for a small latecomer country – the structural transition from the primacy of the humanities and traditional academic science to the dominance of technological R&D has taken place very rapidly, over three to four decades.
- ◇ The transitional processes have increased and intensified the connections and co-operation between universities and other research institutions. The whole R&D system is becoming internally more integrated and dynamic, which also affects the internal life of the previously more separated institutions. This integration tendency is reflected in the novel concept of Triple Helix which refers to a “spiral” development of university-industry-government relations in knowledge production (Triple Helix, 1996). A key issue arising here is to what degree the institutional integration will be based on functional division of labour rather than on eliminating the existing institutional differences (cf. Rappert, 1995).
- ◇ As the relative share of universities’ own budgets of the overall R&D funding has declined, this has made the universities – in order to survive as research institutions – more dependent on various external sources of research funding, both national and international. The universities are therefore actively assuming new kinds of economic functions and “entrepreneurial” activities in order to capitalise on their research services and products (Etzkowitz, 1994).

- ◇ While offering new opportunities for universities to strengthen their institutional positions and research bases, the accommodation of new functions also creates new problems and tensions inside academia. These involve such problems as functional overload (are universities trying to do too many different things at the same time?), the need to renew research organisation and management, and coping with the emerging tensions between old and new cultures of research and research evaluation (Elzinga, 1995; Geiger, 1990).

The evaluation boom: from traditional to new forms of research evaluation

Traditionally, evaluation of research by peers has always been an inherent part of scientific activity and discussion. The new collective and institutional forms of evaluation, so actively developed during the past five to ten years, have largely resulted from external influences and science-policy aspirations. The overall aim has been to redirect the limited, and even decreasing research resources on the basis of science policy objectives and the demands on “accountability” in science. The evaluation boom has raised active public discussion which, however, has mainly concentrated on the technical implementation of evaluation and the indicators used in ranking universities and scientific disciplines.

Since the 1960s, the Finnish science and higher education policies have been institutionally, and also largely by their contents, separated from each other. The Academy of Finland has been responsible for science policy and the Ministry of Education for higher-education policy. Up until now, there have been few active or systematic efforts to integrate these policies. Moreover, even though the primary aim of the Academy of Finland is to finance basic research, there are no coherent links between the Academy and the universities. Recently the Academy and the Ministry have made several efforts to promote the conditions of research and research training together, e.g. by launching the new national programmes of post-graduate education and renewing post-doctoral education.

In Finland, the traditional forms of research evaluation commonly focus on the level of individual scientists, research projects and teams. Evaluation of research can be considered as a continuous process that can be applied at different phases of a research cycle: before (*a priori* evaluation), during (on-going evaluation), and after (*a posteriori* evaluation). Evaluation can also cover all levels of R&D activity, from the macro-level of the national R&D system downward to the various micro-level components. Methods for different levels of evaluation, for different stages, and for different types of research activity have been developed and used with varying degrees of reliability.

Evaluation of research projects or individual researchers has been conducted beforehand, e.g. for all research plans submitted for support to the Academy of Finland. On the project level, evaluations of on-going research are also common, e.g. by monitoring groups of experts appointed for major projects by the Academy of Finland. *A posteriori* evaluations take place on the level of individual researchers when competence and achievements are evaluated, for example in connection with academic appointments (Stolte-Heiskanen and Kaukonen, 1989).

The new collective and institutional modes of evaluating scientific research were actually initiated outside the universities. The scientific body responsible for the evaluation of science has been the Academy of Finland which, already in the 1970s, started the practice of evaluating research priority areas and programmes. A significant change in research evaluations took place when they became a regular feature of official science policy in the beginning of the 1980s. The directives of the Science Policy Council, published in 1981, called for systematic undertaking of evaluations of different R&D sectors (Science Policy Council, 1981). Consequently, the Academy of Finland began (in 1983) to undertake a series of evaluations on certain domains of research and on whole disciplines (described in more detail below).

The new evaluation practices started to influence the everyday life of university-based researchers more profoundly in the beginning of 1990s as the evaluation boom gained new momentum. This was related to new demands on university activities, research included, to be more accountable, efficient and produce “top results” according to international standards. These demands were now reinforced by real threats of cutting university funding¹, on the one hand, and by the incentives of getting extra resources for good performance, on the other.

Quite obviously, the evaluation of science and university activities has come to stay. In a small country like Finland, there is a permanent imbalance between the scarce resources and the potential needs to develop research. The issues of evaluation should therefore be discussed more analytically and not as technical and practical problems only. In order to understand background factors and objectives of evaluations, their development should also be analysed within a broader science-policy context. In this regard the basic concepts of Finnish science policy have undergone major changes during the post-war period as the following brief analysis indicates.

The essential dimensions of science policy

The long-term changes in Finnish science policy may be briefly characterised by selecting some key concepts and characteristics from different periods. In the post-war period of the 1950s, there were both traditional academic and elitist tones in Finnish scientific life, but at that time the emphasis was placed on the national sciences and the “highest cultivation of spirit”. The 1960s were a transition period characterised by the expansion and modernisation of research and higher-education institutions. As the first boom of science policy took place in the early and mid-1970s, the most valuable attribute of scientific research was “social relevance” and the advancement of welfare-state and democratic (social) policy objectives. In the early 1980s, this societal, soft-science orientation was replaced by the new priorities of technology and innovation policy, with a parallel emphasis on the role of basic natural sciences.

An active turn towards internationality in Finnish science policy took place in 1985 along with membership in the Eureka programme. The new orientation culminated in 1989 when Finland decided to apply for full membership in the European research organisations CERN (European Laboratory for Particle Physics) and ESA (European Space Agency). In the 1990s, the internationalisation of Finnish science has been regarded as a key priority, if not the primary one, in science policy. The European Union currently plays a central role in Finnish science and technology policies. After joining the European Union in 1995, Finland has been an active member both in European R&D policy making and in participating in EU programmes. For a newcomer, the main aim seems to have been quite pragmatic, to maximise the research funding coming from EU sources, at which Finland has succeeded well.

The beginning of the 1990s has also witnessed a rapidly growing pressure of accountability on science, following the international tendencies and reinforced by the economic recession and the neo-liberal turn in Finnish economic and social policy (Alestalo, 1993). The policy change is interestingly visible in science-policy terminology: the recent key word of “creative research environments” from the 1988 programme of the Academy of Finland has been replaced by the current buzzword of “accountable research units”. A central aim of the new science policy is to identify the “top units of scientific research” (or centres of excellence), and to compete for international prestige and funding. On the other hand, current science policy is regarded as an inherent part of supporting the national system of innovation in Finland. In line with this, the substantial additional funding for R&D in 1997-99 (to which the Cabinet

1. Considerable cuts in resources actually took place during the economic recession of 1991-95 as the basic budgetary funding of universities decreased by 15 per cent.

has committed itself and which should increase the research funding to 2.9 per cent of GDP) will be allocated mainly to the applied, technology-intensive fields. The extra funding will be channelled in the first place through Tekes (Technology Development Centre) and the Academy of Finland, and to a relatively lesser extent, the universities and sectoral research units.

The above schematical analysis of the changes in Finnish science policy reveals some basic dimensions in the development of national R&D systems, especially from a small country perspective. It may also help to understand the broader context and reasons for the development and present boom of research evaluation. Based on the analysis of the Finnish development, the following five science-policy dimensions that are relevant for the evaluation of research are distinguished below.

- ◇ *National vs. international orientation.* In Finland, a clear shift from national to international orientation has taken place in science policy in the long run. This shift has not been linear however; the national “closure” in science policy was most evident in the 1970s. The international orientation of science and its shifting geographical directions have reflected changes in geopolitics and national foreign policies.
- ◇ *Soft vs. hard S&T.* The transition from the domination of the humanities (1950s) through social sciences (1970s) to the natural and technical sciences (1980s and 1990s) has been evident. At the same time, the concept of science policy has undergone similar changes - from “soft” science policies to ‘hard’ science, technology and innovation policies. From the point of view of research evaluation, it is important to note that the ideal model and criteria of science have changed as well. The concept of science has so far remained rather monolithic - one model has more or less replaced the other. The time of a pluralistic or synthetic understanding of science and science policy is not yet visible.
- ◇ *Narrow excellence vs. broad competence.* This is a strategic choice for a small country: should science policy be highly selective and aim at reaching the international top in some specialised areas, or should it support the development of broad national competence in the first place? In this regard, post-war Finnish science can be characterised as a nationally-oriented elitism (the old Academy of Finland), whereas the science policy of the 1960s and 1970s strongly emphasized the broadening of the science system and its regional base in the spirit of equal opportunity. The 1980s have witnessed a more selective policy, especially in favour of technological innovation policy, and most recently, in favour of scientific excellence as measured by international criteria.
- ◇ *Academic basic research vs. societal utilisation.* In the long term, one may note a distinct shift of emphasis from the academic, “curiosity oriented” research toward an emphasis on societal utility and application of research, or from basic to applied research (cf. Gibbons *et al.*, 1994). The societal utility or accountability of research, however, has been understood differently in different times: in the 1970s the focus was on societal and public policy relevance of research, in the 1980s the argumentation based on technology and innovation policy objectives gradually emerged at the forefront, while the early 1990s have been characterised by a more general market-oriented accountability thinking. In understanding the concept “societal”, this has meant a transition from the public to the private sphere. In basic research a certain renaissance took place in the early 1980s, but it mainly concerned basic research in natural sciences.

- ◇ *Intrascientific competition vs. mutual co-operation.* In this respect it is difficult to discern any clear long term trends or changes. Both elements are important and always present in scientific work and they need not be opposites that exclude each other. Science policies can influence the degree and interrelationship of competition and co-operation in science, for example by emphasizing respective criteria of evaluation and funding. However, this issue has remained rather marginal in Finnish science policy until the 1990s. Perhaps it is so that only external pressures and promises on funding make researchers and institutions compete and/or co-operate, depending on the policy guidelines! Recent science policy in Finland has advanced more competition than scientific networking and co-operation. Quite recently, however, the Ministry of Education also started to emphasize the co-operative aspects while, for instance, making decisions on the new graduate (doctoral) schools in Finnish universities. It may be that participation in the European R&D programmes will have similar effects in the future.

The above analysis of the turns in science policy development shows that science itself is far from being monolithic and the evaluation of research involves multiple dimensions. The review also shows that Finnish science policy has undergone several turns at the expense of continuity as it has been most sensitive to economic and political fluctuations, especially as concerns the economic recessions of the mid 1970s and early 1990s.

Research evaluation for what?

Different potential interests become apparent when we try to analyse at the background factors and objectives of doing research evaluation. These objectives are not always clearly pronounced in the public and they may even serve conflicting interests. Following is a brief and tentative look at the driving forces and objectives of research evaluation from the Finnish perspective.

In Finland, the traditional academic self-evaluation practice remained intact until the late 1960s. New elements started to emerge in the early 1970s when the new-born Academy of Finland launched its science-policy priority programmes. For some years, the societal relevance of research was strongly emphasized in the funding decisions made by the Academy. However, the heated public disputes over some politically sensitive projects and, finally, the economic recession put an end to this evaluation experiment.

When the new kind of disciplinary evaluations were started by the Academy of Finland in the early 1980s, their character and background were quite different from the soft-science oriented experiment of the 1970s. The new evaluation practice helped to strengthen the position of the basic natural sciences (a “back to basics” movement), on the one hand, and to support the emerging technological innovation policy, on the other. At the same time, as the evaluations by international experts revealed some problems in the disciplinary research practice (for example, inadequate co-ordination and international mobility), they also acknowledged that these fields in general were high on an international level. The disciplinary evaluations thus indirectly created a new standard and model of good science and its main characteristics.

The most recent developments of the early 1990s, or the “evaluation boom”, are closely related to external factors, such as the deep economic recession, cuts in R&D funding and the strengthening of the new market-oriented policy of accountability and competitiveness. From a political and administrative perspective, it was more convenient to reallocate decreasing funds through quality competition than to cut them. Research evaluation is gradually becoming a legitimate tool to (re)allocate resources so that the research community itself is more or less involved in the process.

When we compare the effects of the two economic recessions (mid-1970s and early 1990s) on science-policy activities, an interesting difference may be observed. In Finland, the former recession resulted in a deep break in science policy; it practically stopped the visible science-policy activities for several years. Now quite the opposite seems to be true. Science and higher-education policies have become more active than ever since the university reforms of the early 1970s. How can this difference be explained? A tentative list of explanations would include: the overall growth of accountability pressures in public administration, the functional diversification of universities toward “multiversities” which has decreased the earlier academic opposition to external influences, and the availability of new tools (methods) and international experience of university regulation and evaluation.

If, finally, we look at the background interests of carrying out research evaluation more analytically, one may distinguish the following five reasons for evaluation.

1. The qualitative and quantitative evaluation of science may be used as a normal, inherent part of developing research activities (which need not be limited to self-evaluation only).
2. Evaluations may be used as a tool in science-policy decision making to define priorities and to decide on preferential funding.
3. Evaluation may aim at strengthening the administrative and political legitimacy of the research system, or some part of it, by showing its efficiency and accountability to relevant audiences (cf. the notion of “symbolic evaluation” by Foss-Hansen, 1995).
4. Evaluations may be used as a tool in intramural competition between scientific fields in which the dominating fields can influence the criteria and thus gain competitive advantage relative to others.
5. Evaluations can be used as a seemingly neutral means of allocating resources to and within the science budget by the governing bodies. Here, the problem may be that in the name of research evaluation, socio-political choices between different R&D policy sectors are being made.

Obviously, the reasons for carrying out research evaluation can differ considerably and also involve potential contradictions between the various objectives of evaluation, be they explicit or implicit. From the point of view of maintaining the qualitative diversity of science, the often less explicit objectives four and five (above) are the most problematic. If there is only one category for the internationally-oriented science and technology in the evaluation exercise, the other fields like social sciences and humanities are doomed to be the losers. Therefore, it would be important to discuss the objectives and criteria of research evaluation more openly and critically than has been done so far.

Current actors and practices of research evaluation

The main nation-wide activities which currently involve active research evaluation in Finland are described below in more detail. The presentation is not meant to be exhaustive, however. In addition, mention could be made of the recent structural development programme of universities as well as of intra-university research policies where evaluative data have been used in defining research strategies and priorities.

Disciplinary evaluations by the Academy of Finland

When started in the mid-1980s, the main thrust in the disciplinary evaluations was on the basic (natural) sciences, but later on, the sphere of research domains evaluated was expanded and became more versatile. Some examples of the 18 scientific disciplines evaluated up to the present are: inorganic chemistry (1983), experimental nuclear and high energy physics (1985), automation technology (1986), environmental toxicology (1988), peace research (1990), research on climate change (1992), legal science (1994), and research on molecular biology and biotechnology (1997). In addition to disciplines, the evaluations have covered two research institutes (national health; low temperature physics), two research programmes (climate change; materials and structural research), and Finland's membership in the International Institute for System Analysis (IISA).

The evaluations have relied heavily on the classical peer review. However, where the national scientific community is relatively closed and small – as in Finland – the well known problems of objectivity associated with (domestic) peer review become magnified. The solution adopted is to draw peers from the international community. Thus, almost all the evaluations have been conducted by an international evaluation group. The major exception has been the evaluation of the science of education where the choice of experts was limited to Finnish researchers only (also, in the field of legal science, most of the panel members were Finnish).

The disciplinary evaluations have emphasized the assessment of research excellence in terms of international prestige and contribution to the forefront of science. It has become obvious, though, that the reliance on international experts in evaluations has limited potential in the applied and social sciences and humanities. In these fields – publications are mostly in the native language – there are no necessarily competent foreign peers, and a variety of criteria of assessment are needed.

The use of international experts may not be unproblematic even in relatively uncontroversial research areas, since foreign peers are usually confronted with the current international state of affairs while lacking knowledge about the historical, structural and organisational context of research activities in Finland. An equally important issue is that the use of international experts is explicitly connected with the criteria of assessment in terms of “comparison to the international top level”. The relative position of a field in this regard, if taken as an exclusive criterion of evaluation, is only applicable to a few non-controversial basic research areas.

The standard procedure in these evaluations is that an evaluation report is prepared collectively by invited foreign experts, based on summaries of research activities and publications of research groups and on-site visits and interviews with scientists. With some differences, the assignments of the evaluation groups have consisted of the evaluation of the sufficiency and appropriateness of research posts, equipment and other resources, the quality of research and the future plans for research development. Depending on the substance, the research conducted in Finland has also been compared with corresponding research conducted in other countries. All the disciplinary evaluations have also included evaluation of post-graduate education (training of researchers) in respective fields. The data gathered for the evaluation have consisted of scientific publications, statistics of higher education and research, general overviews of the research organisation, its staff and activities, funding profiles and descriptions of future plans. The evaluation reports also involve detailed descriptions of individual departments. The Academy of Finland has covered the evaluation costs which, on average, have amounted to Mk 300 000 (~ECU 60 000) per evaluation.

In the current performance contract with the Ministry of Education, the Academy of Finland has made an ambitious commitment to evaluate the state of Finnish science in all university-based research fields in

two-three years. This effort will include the continuation of disciplinary evaluations in selected research fields. In addition, last year the Academy started the preparation of state-of-the-art reviews which cover all scientific fields and will eventually be summarised at the level of the four Research Councils.

From the point of view of the universities, it seems that the disciplinary evaluations undertaken by the Academy have remained rather external, but also uncontroversial, activities as their effects have been felt and discussed mainly within the specific fields concerned (with the exception of the first evaluation on inorganic chemistry whose results were widely debated by the public).

The policies for promoting centres of excellence in research

In 1993, the idea of nominating centres of excellence in research (or “top research units” in Finnish language) was introduced in the science-policy outlines of the Academy of Finland and the State Council (18 June 1993). The new policy can be seen as providing a strategic means for promoting the internationalisation of the Finnish science system. By channelling funding to selected top units, it is believed, the possibilities for Finnish scientists in the competition for international research funding will be increased. Moreover, the centres-of-excellence policy is seen as a means of strengthening political and public confidence in science.

The Academy of Finland and the Council for Higher Education (Council for Higher Education Evaluation, since 1996) and the Ministry of Education have been responsible for the selection process. The first nomination process was co-ordinated by the Council for Higher Education. The first nomination in 1993 had to be accomplished in two weeks by the commission of the Ministry. In the next round of nominations (1994), the Academy of Finland made the primary and actual selection. In the third round (1995), the Academy proposed to the Ministry that no new centres of excellence should be nominated that year. So far, a total of 17 research units or departments have been nominated and have received extra funding. Recently, the Academy of Finland has devised a plan to triple the amount of research performance-based funding (Alestalo and Tuunainen, 1996). The Academy has also decided to transfer a part of its former junior-researcher posts to the centres of excellence as post-doctoral posts. A working group of the Academy is currently preparing the guidelines on how to continue and develop the top-units policy from now on.

Research performance-based budgetary funding

Currently, approximately 90-92 per cent of the budgetary allocation to universities is based on the number of masters and doctoral degrees produced at the universities. The doctoral degrees (32 per cent), are regarded as indirectly reflecting the university’s research performance. Additionally, 3-5 per cent of the basic budget is allocated according to quantitative and qualitative performance indicators, including the centres of excellence, funding from the Academy of Finland, international funding, international exchange of teachers and researchers, graduate employment, and estimation of the renovations and development trends at university.

The rest of the basic budgetary funding, 3-5 per cent, is reserved for specific project funding in the nationally-defined priority areas (including both research and educational projects) which are agreed on in the performance negotiations with the Ministry. The meaning of this sharing is to give the Ministry of Education the possibility of launching projects and supporting activities which are considered to be of nation-wide importance, and to give the universities a possibility for “new openings” in research, teaching and development work. At the moment, some examples of the largest projects are the Graduate (doctoral) Schools and projects related to the advancement of the Information Society, both in research and teaching.

As a whole, one may estimate that direct research performance-based funding currently accounts for 3 per cent of the universities' basic budgets. With indirect funding on the basis of doctoral degrees included, the share amounts to *circa* 35 per cent.

The budgetary evaluations are mainly based on the information contained in the national KOTA database (the name is an acronym of the committee on whose work the database is based). KOTA is a statistical database containing data which describe university performance by institutions and by educational fields since 1981. The database was established to give access to up-to-date aggregate data for higher-education and research planning, monitoring and evaluation. At the moment, KOTA contains primarily educational data, but also information on university staff, appropriations, international exchange in research and scientific publication.

In 1995-1996, as a part of the development of result-based funding and management, the Ministry of Education nominated two working groups to outline the future trends and needs for developing the funding system. As the universities were constantly concerned about "forgetting the research funding" in the calculative model, the Ministry nominated a special group to investigate the possibilities for including research performance in the overall funding mechanism. Quite recently, the working group suggested in its final report that 35 per cent of operational funding should be allocated on the basis of the results of research conducted in the universities; in addition, 15 per cent should be allocated on the basis of doctoral degrees. The evaluation of research would be carried out on all university departments and research units in all research fields by the Academy of Finland every third year (by using the peer review method). The group suggested, by following the British model, that all units be ranked on a scale from one (lowest) to five (highest). The total funding allocation of a university would be the sum of all disciplines represented in the university.

The proposal, however, was rather heavily criticised by the universities and other relevant parties. The proposed evaluation procedure was seen as problematic and the universities were unanimous in their opinion that the proposal would give overwhelming science policy power to the Academy. In addition, it was argued that the new system would drastically change the allocation of funding between universities, and would erect normative standards for profiling resources among units. The Ministry has currently "frozen" the initiative, due to the criticism.

On the effects of evaluation

As could be expected, follow-up studies have shown that the reactions of the scientific communities concerned and individual researchers to disciplinary evaluations have been generally enthusiastic in cases where the assessment has been positive, and critical of the reliability of the approach when the conclusions were negative (see Luukkonen and Ståhle, 1993). In more general terms, one may estimate that the evaluations have, by raising awareness of the key criteria used, reinforced the international orientation of researchers, as well as that of their publishing behaviour. It also seems that the relatively high age at which one obtains one's doctorate in Finland, which has been criticised in several evaluation reports, has lowered over the last few years. There are, of course, other factors involved here and it is difficult to assess what has been the specific impact of evaluations, but they certainly raised the issue to the science-policy agenda.

Issues and problems involved in research evaluation

As mentioned earlier, the evaluation of science and university activities is here to stay. For this reason alone it is important to notice that the evaluation of research is not just a technical or practical problem. Rather, it is a complex, multi-dimensional issue which should be analysed and discussed more analytically than has been the case so far.

The most essential questions might include the following:

- ◇ What are (have been) the basic science-policy arguments and objectives of evaluation?
- ◇ What effects - explicit and implicit, direct and indirect - do the evaluations have?
- ◇ How does one take into account the diversity and qualitative differences of disciplines and research fields?
- ◇ What is actually meant by “accountability”? Who should be the beneficiaries of research and who should evaluate its results?
- ◇ How can the benefits of evaluation be maximised and the negative consequences minimised?
- ◇ How much time and money is evaluation actually worth?
- ◇ Finally, a question remains concerning the evaluation of evaluation, as the researchers have the right to ask what kind of evaluation is reasonable and most beneficial from the point of view of research itself.
- ◇ What would an “ideal” evaluation look like, and how do the perceptions of various parties differ in this respect?

The first two points have already been discussed above, and some of the others will be briefly commented on below.

Over the last few years, evaluation activities have become an inherent part of university management. Evaluation has become an administrative imperative. To survive in the struggle for scarce resources, every academic unit and researcher has to play the game and play it according to prescribed rules. *Administrative evaluations* involve more control than the traditional ones, as they tend to emphasize the supposedly commensurable and quantifiable aspects of research performance, an approach which may greatly differ from the traditional approach of qualitative self-evaluation by the research community and, hence, produce controversial and even counter-productive results, e.g. publishing behaviour may become standardized – according to certain criteria – despite the disciplinary differences.

Overly administrative evaluations become easily defensive, self-manipulating and rhetorical. Scientific units soon learn to use the right words, like “top units” and “international spearhead”, in defining their objectives and in marketing their products. The university and science-policy administration will thus face the “post-modernist” dilemma of epistemic ambiguity – what is “real” and what is rhetorical, or is there any practical difference between the two anymore? Aant Elzinga (1995) has observed a somehow related development in scientific norms and values which he describes as a transition from traditional epistemic values to a more complex social epistemology where “credibility” and “trust” are becoming the new key words.

Still, research evaluation is a very sensitive matter, at least from the point of view of the individual researcher. Historical evidence shows that incompetent evaluation and excessive *administrative control* may easily weaken the scientific motivation and innovativeness needed in research work. If the evaluation

system does meet the qualitative specifics of the field or if it is otherwise felt to be inappropriate, the most active researchers may end up in a contradictory situation: how and to what degree does one reconcile personal scientific interests with the official evaluation criteria? Should one hunt for evaluation points or for new ideas in the first place, or try to do both? This kind of double life on the part of active researchers was a common, unofficial practice in the academic fields of the former socialist countries (Kaukonen, 1994). Of course, the new university evaluation systems do not equal the former socialist plans, but similar effects may still appear. In order to eliminate the problem of double standards and manipulative effects in evaluation, it is therefore important to emphasize the active role of the research community, both in defining criteria for scientific quality and in carrying out appropriate self-evaluations.

Balancing and integrating the two relatively isolated “*evaluation cultures*”, the one within the scientific community (internal self-evaluation) and the other, the official and administrative evaluation programmes (external evaluation), may be problematic. Can evaluations simultaneously serve two masters whose interests and needs for information seem to differ in several respects, e.g. common quantitative indicators vs. qualitative assessments, or should the evaluations be different for different purposes? In any case, there should be a constructive, and hopefully, effective dialogue between the two cultures.

Another problem in the evaluation discussion and practice concerns the very *concept of science*. In Finland, science has been understood as a relatively monolithic phenomenon – not paying due attention to the *disciplinary diversity of science*. The rather unproblematically-taken “ideal type” of good research practice clearly has its origins in basic natural science and medicine, while now also covering the new high-technology fields. In the 1960s and 1970s, Finnish sociology was accused of “scientific imperialism” as it tended to dominate other fields across its boundaries (Heiskanen, 1982). Much in the same vein, the natural scientific, or technoscientific, model has set a general standard for the evaluation of science in the 1980s and 1990s. This is perhaps most visible in how the role of internationality of science is emphasized and understood in the evaluations. It seems that among the strong fields of natural science and medicine, it is very difficult to understand that there can be real and “natural” differences in the national and international orientations between disciplines, or that the emphasis on national aspects in research does not need to imply scientific “underdevelopment”.

A key problem of research evaluation concerns *the notion of scientific quality*. Every researcher naturally tries to accomplish something qualitatively new and important in his/her speciality. This seems to be the only behaviour common to all fields of science. The concrete criteria of quality of research, or what is regarded as good research and what is not, differ greatly among different scientific fields, depending on substance and theoretical and practical purposes of research. Also, the basic characteristics of research practice, such as the individual vs. collective character of research work, and the needs for collaboration, vary markedly. In general, however, the empirical studies (Hemlin, 1991; Kaukonen, 1997b) indicate that it is possible to define some basic dimensions and attributes of scientific quality which, however, have different weight (cf. internationality) and concrete contents (cf. societal relevance and utility) in various disciplines. Quality is diversified, too, and this basic fact should be recognised in science policy and research evaluation.

For instance, the great variety in *publishing behaviour and citation practices* may be largely explained on the basis of inherent disciplinary differences. This indicates that the often-used bibliometric indicators have low validity in comparing research performance between different kinds of disciplines. Science in small countries is faced with the decision on whether or not to publish in national or international publication channels. Publications in international fora may be an appropriate indicator of participation in international advancement of knowledge in relevant areas of research. In other fields, however, the spread of research results through national channels may be more congruent with research goals. This conclusion is supported by an ongoing study on the Finnish scientific elite covering all major scientific fields. The

data clearly demonstrates the diversity of disciplines and their “best research practices” as regards the national *vis à vis* international orientation, and many other aspects as well (Alestalo and Kaukonen, 1995).

Taking into account the disciplinary differences, the uncritical use of the currently popular *concept of “international excellence”*, or its Finnish version “top research”, also becomes problematic. Basically, the problem concerns the possibility of defining what constitutes the international top of research in various fields, and to what degree a consensus prevails on this definition among researchers. Our Finnish data indicates that the definition of scientific excellence, or the best research groups in the world, is feasible if the research field is universal enough, has a relatively sound paradigmatic basis and a well-articulated international research agenda, and if the field is specified narrowly enough. This conditionality implies, however, that the general science-policy relevance of defining centres of excellence necessarily becomes quite limited. In nationally or locally-oriented research, for instance, the search for the absolute international top would lead to paradoxical results.

Especially in small countries like Finland, where a considerable share of scientific activities is necessarily related to building up endogenous capabilities, more intrinsic criteria for assessing the value of research are also needed. These include such factors as, e.g. the scope of research, contributions to producing new syntheses, flexibility and adaptability to problems of specific national interest, promotion of transdisciplinary capabilities to ensure innovative potential, etc. – not to mention the criteria of societal relevance.

Therefore, in the evaluation of science it is important, but not easy, to take the *inherent diversity and multidimensionality of the research system* into account. For instance, questions about quality, internationality and social accountability of science are not self-evident, but complicated issues which need to be studied. Accordingly, in the evaluation of scientific performance and quality, this *diverse ecology of science* should be accepted as a basic point of departure. Here one may draw parallels between ecology and environmental policy. I believe that taking the innate diversity of science seriously, in the same way as biodiversity has been gradually accepted in ecology, would be a major step forward. Instead of being seen as a harmful deviance from the norm, the diversity of scientific fields could be understood as a valuable resource which may generate new ideas and innovative connections. The positive role of diversity has recently come up also in such areas as journalism and industrial high-tech development (Johnson, 1992) where diversity is actually seen as a guarantee of qualitative and innovative development.

In more practical terms, one should also ask *how much time, money and effort* should be reasonably used for evaluations and what kind of evaluation practices would be optimal in this respect? As everybody knows, evaluations are costly and time-consuming activities, whose justification beyond the value of a purely academic, or bureaucratic, exercise is that they serve some useful purpose, i.e. they have some impact on something. What would be the actual cost/benefit ratio of evaluation activities if all the factors and effects involved were taken into account? Some university colleagues already speak of “evaluation exhaustion” or “fatigue” as the time remaining for the main activities (research and teaching) is becoming all the more scarce. This situation, of course, will increase in other secondary activities as well.

Finally, the evaluation issues concern the very *concept of university*. Traditionally, universities have been conceived of as self-monitoring, autonomous systems which are now being confronted with the idea (or fear) of universities as externally-controlled or regulated organisations. In my view, this abstract dualism is out-of-date and not useful in analysing the actual development. Rather, universities are developing towards discursive, network-like institutions which try to accommodate and fulfil a growing variety of tasks. In the area of research, these range from traditional “curiosity-oriented” basic research to “customer-oriented” research services. This implies a growing pluralism and decentralisation of research activities and, hence, a need for discursive co-ordination and evaluation of research.

Conclusion: The ideal evaluation?

Although there is a growing consensus on the need for R&D evaluations, there is less agreement about the relevance and adequacy of existing evaluation approaches. Differences in goals, methods and objects of evaluation notwithstanding, on the whole, evaluation studies have been more product than process-oriented. The methodological emphasis has been on the development of indicators and procedures for assessment of the current state or past achievements of the objects of evaluation. However, the processes and material conditions affecting the state of affairs are equally, if not more, important in order to arrive at balanced and justified judgements. A broader view would be important, particularly concerning the evaluation of (basic) research at universities which, at least in Finland, have struggled over the last few years with the often conflicting pressures of decreasing resources and new demands for accountability and functional development (i.e. new functions and tasks assigned to universities).

To conclude, I would like to present some general points which I personally regard as important for a positive or “ideal” integration of evaluation into research practices.

- ◇ *First*, the evaluation should be a part of normal scientific activity and its development in the first place, to ensure that the specific characteristics and objectives of the research fields are taken into account. Instead of administrative control, the evaluation should provide positive incentives and motivation for research work – it should be a secondary, “research-friendly” activity.
- ◇ *Second*, in order to minimise the negative effects and the manipulative use of evaluation, its criteria should be based on a pluralistic concept and understanding of science.
- ◇ *Third*, a comprehensive evaluation of scientific results, or accountability, presupposes both active self-evaluation, or self-reflection, and broad hearing of external views when relevant; in assessing the societal relevance and impact of research, the citizens’ perspective should also be involved.
- ◇ *Fourth*, instead of quantification and commensurability of the indicators, the evaluations should, in the first place, provide qualitative information of the objectives, state and results of research. The use of numbers, when needed, should build on qualitative accounts in order to make the numbers meaningful. The transparency of evaluations is important here.
- ◇ *Fifth*, as the use of time is becoming an increasingly critical factor in universities, the researchers should have the right to concentrate on actual research work and to minimise other activities. Therefore, unnecessary administrative functions, also related to evaluation, should be eliminated.
- ◇ *Sixth*, in developing reasonable forms and ways of research evaluation, science studies would have an important contribution to offer.

REFERENCES

- ACADEMY OF FINLAND (1988), *Science policy program*.
- ALESTALO, M. (1993), "The rise of neo-liberalism in Finland: From the politics of equal opportunity to the search for scientific excellence", *Science Studies* 6, 2, pp. 35-47.
- ALESTALO, M. and E. KAUKONEN (1995), "Reaching for the international forefront of science", manuscript based on the paper presented at the Conference on Dynamics of Science and Technology, University of Tampere, 8-9 September.
- ALESTALO, M. and J. TUUNAINEN (1996), "Evaluation at the University of Helsinki: Project EVALUE, Programme TSER", European Commission, DG XII.
- ELZINGA, A. (1995), "The historical transformation of science with special reference to 'Epistemic drift'", manuscript based on the paper presented at the Conference on Dynamics of Science and Technology, University of Tampere, 8-9 September.
- ETZKOWITZ, H. (1994), "Academic-industry relations: a sociological paradigm for economic development", in L. Leydesdorff *et al.*, eds., *Evolutionary economics and chaos theory*, pp. 139-151, Pinter Publishers.
- FOSS-HANSEN, H. (1995), "Organising for quality - a discussion of different evaluation methods as means for improving quality in research", *Science Studies* 8, 1, pp. 36-43.
- GEIGER, R. (1990), "Organised research units - their role in the development of university research", *Journal of Higher Education* 61, 1, pp. 1-19.
- GIBBONS, M. *et al.* (1994), *The New Production of Knowledge*, Sage Publications.
- HEISKANEN, I. (1982), "Yhteiskuntatieteet, käytännön yhteiskuntateoria ja maamme älyllinen ilmasto (Social sciences, practical social theory and the intellectual climate of our country)", *Helsingin yliopiston valtio-opin laitoksen tutkimuksia*, sarja A:59.
- HEMLIN, S. (1991), "Quality in science: Researchers' conceptions and judgements", Department of psychology, University of Gothenburg.
- JOHNSON, B. (1992), "Institutional learning" in LUNDVALL, B.-Å., ed., *National systems of innovation*, pp. 23-44, Pinter Publishers, London.
- KAUKONEN, E. (1994), "Science and technology in Russia: collapse or new dynamics?", *Science Studies* 7, 2, pp. 23-36.

- KAUKONEN, E. (forthcoming, 1997a), "Science systems in transition: problems and perspectives", in D. Gibson, ed., *The Science City in a Global Context*, The University of Texas at Austin, JIMT-Program Monograph Series.
- KAUKONEN, E. (forthcoming, 1997b), "Science policy, research evaluation and the diversity of science: a discussion based on Finnish experience", in M. Hyvärinen, ed., *The Institutes We Live By*, RISS, The University of Tampere.
- LUUKKONEN, T. and B. STÅHLE (1993), "Evaluation of research fields: scientists' views", Nordic Council of Ministers, Nord 15.
- RAPPERT, B (1995), "Shifting notions of accountability in public- and private-sector research in the UK: some general concerns", *Science and Public Policy* 22, 6, pp. 383-390.
- SCIENCE POLICY COUNCIL OF FINLAND (1981), *Science policy program*.
- STOLTE-HEISKANEN, V. and E. KAUKONEN (1989), "Trends and problems in Finnish evaluations of science", in A.G. Kharchev and J-P. Roos, eds., *Sociology and society in Finland and Soviet Union*, The Committee for Scientific and Technological Co-operation 29, Helsinki, s., pp. 143-154.
- TRIPLE HELIX: UNIVERSITIES AND THE GLOBAL KNOWLEDGE ECONOMY (1996), "A Triple Helix of University-Industry-Government Relations", book of abstracts of the conference held in Amsterdam, 3-6 January.

CHAPTER 2. EVALUATION OF SCIENTIFIC RESEARCH IN THE NETHERLANDS

Marcel A.M. Eiffinger, Ministry of Education, the Netherlands

Executive summary

This paper takes stock of Dutch evaluation practice for basic and strategic research. The section entitled “Mapping Dutch evaluation practice: past and present” characterises three main types of assessment: evaluation of scientific research institutes and groups; programme and project evaluations; as well as sectoral and societal assessment of research. “Looking ahead” deals with the policy questions of how to link evaluation and funding, and how to bridge evaluation practices.

Dutch research evaluation practice took off in the mid-seventies. Since then, a rich evaluation culture has slowly emerged. Gradually players from the whole spectrum of scientific research have become involved, and research evaluation has become an integral element of Dutch science policy. Not because policy makers thought evaluations important in their own right, but because they considered them a useful instrument for (occasionally) restructuring the science system. The principle of “governance at arm’s length”, reflecting autonomous decision-making by R&D institutions, came to be a central principle underpinning evaluation policy, particularly with regard to institutional evaluations. Rather than carrying out *evaluations of research institutes and groups* on its own, Dutch government has confined itself to triggering evaluations, to be performed, or at least commissioned by, the research organisations themselves. The typical model for such evaluations is one of external evaluations usually undertaken by independent peers, often preceded by internal or self-evaluations. Sometimes peer reviews are backed up by bibliometric analyses. Research organisations themselves are held responsible for initiating or carrying out these evaluations. They are expected to orient themselves on their environment and to formulate a research and institutional strategy, based on perceived strengths and weaknesses, and laid down in strategic plans. Thus quality assessments are used as an important expedient for defining research strategy; and not so much as an input for allocative decision-making, although this may change in the future.

In contrast with this, *evaluations of research programmes and projects*, e.g. the strategic programmes that seek to create or strengthen research nuclei in areas of (potential) scientific or societal relevance - have been used for financial decision-making concerning the launching or continuation of these programmes and projects. Initially programmes seem to have been evaluated “in house” and in a rather informal way by the responsible agency. Later on external and more formal review became the vogue. As a rule these programmes show a pragmatic mix of expert evaluation backed up by systematic data gathering and encompass both an assessment of the scientific quality and “impact” evaluation with regard to the quality of research from a societal perspective.

In the course of time - as is illustrated by the goals of these programmes - more and more attention has come to be paid to the *assessment* of research in specific sectors and to the relevance of research from a societal perspective. This trend originated in the mid-seventies and asked for new types of research and research assessments, as well as new fora to discuss the societal research priorities. Thus new actors, bodies and research evaluation practices could be identified, such as the Sector Councils. These councils,

established in the late seventies for a small number of areas, are still in existence and function on the basis of tripartite discussion between those involved in research, the users of research and policy makers. Their task is to “steer” research in areas of societal importance. To underpin their advisory tasks, the councils can also initiate evaluations. Because of the need for a continuous process of research foresight exercises, and the definition of research priorities on the national level, the Foresight Steering Committee (OCV) was established in 1992. This independent Committee was asked to assist the minister in the formulation and implementation of strategic choices and to develop a systematic long-term focus for priority setting in scientific research.

Although new evaluation practices are called for, for instance because of the launch of top technological institutes and top research schools, evaluation policy in the near future will focus primarily on bridging the various evaluation practices.

Introduction

The context

In the Netherlands a rich evaluation culture has gradually evolved, resulting in a “patchwork system” (Rip and Van der Meulen, 1995). Since the policy for the evaluation of scientific research took off in the mid-seventies and evaluation activities evolved, players from the whole spectrum of scientific research have become involved. The emergence of the intricate web of assessments has been closely interwoven with changes in the science system and the development of science policy, e.g.:

- ◇ changes in the organisation and orientation of research that reflect varying patterns of co-operation - such as the shift from mono-disciplinary to multi-disciplinary research, partly due to a more profound orientation towards the societal relevance of research - called for new types of evaluation;
- ◇ the trend towards decentralised decision-making increased the need for external accountability;
- ◇ budget cuts made strategic choices crucial and provoked self-profiling of universities based upon assessment of strengths and weaknesses;
- ◇ fairly recent policy initiatives, like the creation of the Dutch top technological institutes and top research schools, require additional research assessment practices.

Therefore, the process of enrichment can not be attributed to a single player, specific event or policy impulse.

The Dutch public science system comprises, apart from its 13 universities:

- ◇ The Netherlands Organisation for Scientific Research (NWO), a national research council that funds research projects and programmes mainly at universities as well as a number of NWO-institutes for fundamental research.
- ◇ The Royal Netherlands Academy of Arts and Sciences (KNAW), not only a forum for scientists and scholars, but also an advisory body to the government. It runs a number of fundamental research institutes.
- ◇ The Netherlands Organisation for Applied Scientific Research (TNO), one of the largest organisations for applied research in Europe.

- ◇ The Agricultural Research Service, operating a number of applied research institutes and stations.
- ◇ A number of institutes for fundamental and applied research, or the so-called GTIs.

Complementary to the institutional framework, quite a number of research programmes, such as the Incentive Programmes and the Innovation Oriented Research Programmes, are carried out, promoting not only specific areas of research, but also co-operation between universities and institutes.

In comparison to the OECD-countries in general, the government allocates rather substantial funds as a percentage of the GNP to the universities and institutes. University research in particular is fairly extensive.

Throughout the system, and especially in the sphere of fundamental and strategic research, a multitude of evaluation processes have developed, not only of projects and programmes, but also of institutes and university groups.

The evaluation of disciplinary research and research in specific areas of societal importance is quite extensive. These processes are related to research programming and priority setting as well as foresight exercises.

The paper

It is in the context sketched above that this paper tries to take stock of and characterise Dutch research evaluation practice. In the following section, we will try to characterise Dutch practice by discussing the various types of assessments – such as the evaluation of research institutes and research programmes – from both a historical and a science policy perspective. Readers interested in technical aspects of the methodology of evaluations are referred to the separate boxes, depicting the evaluation criteria applied and the organisational set-up of evaluation practices.

We will conclude this paper by looking ahead. The Dutch “system” of research evaluation is evolving continuously. This calls for co-ordination - both on the level of content and process - so as to avoid a multitude of scattered practices as well as duplication resulting in “assessment-fatigue”; not to mention the necessity to balance the needs for (externally oriented) accountability and (internally oriented) research improvement that underpin most evaluation efforts. The section entitled “Looking ahead” deals with these system dynamics and some policy questions that are arising.

Mapping Dutch evaluation practice: past and present¹

Historical overview

In the Netherlands, evaluation in general had been advocated for some time by such bodies as the General Accounting Office and the Finance Ministry. However, evaluation of research and of policies for science did not take off before the mid-seventies - that is, apart from evaluation of project proposals by ZWO, the predecessor of the Netherlands Organisation for Scientific Research (NWO), the national research council.

1. Much of this chapter builds on the article of Rip and Van der Meulen (1995), *The patchwork of the Dutch evaluation system*, Research Evaluation, April.

As time went by research evaluation gradually became an integral element of the Dutch Science policy. As Rip and Van der Meulen pointed out, not because policy makers thought these evaluation efforts important in their own right, but because they perceived evaluation policy - and evaluations of research institutes and groups in particular - a useful instrument for (occasionally) restructuring the science system. This may help to explain the ad hoc elements in the policy for institutional evaluation that are – to some extent - even nowadays recognisable.

Policy makers conceived of evaluation of research organisations primarily as a matter of “good housekeeping” of research institutions, rather than as a basis for allocating research funds or assessing goal achievement. In accordance with this policy perception, the principle of “governance at arm’s length”, reflecting autonomous decision-making by R&D institutions, came to be a central principle underpinning evaluation policy, particularly with regard to institutional evaluations. Research organisations themselves are held responsible for initiating or carrying out these evaluations. They are expected to orient themselves on their environment and to formulate a research and institutional strategy, based on perceived strengths and weaknesses and laid down in strategic plans. This is illustrated by the Act on NWO and the Act on the Netherlands Organisation for Applied Scientific Research (TON). Both acts prescribe the formulation of such a strategic plan.

At the same time research organisations themselves have also become aware of the need for evaluating the quality of their research and the role these evaluations could play in defining the research and institute’s strategy - partly as a result of budget cuts, particularly from the mid eighties onwards, and because of changes in the organisation of research and institutional innovations - although this awareness was also due to the Dutch Government’s science policy.

In this way evaluations have gradually become an input for (often bilateral) discussions of research organisations with the minister of Education, Culture and Science on the implementation of their strategic plans. The attention paid nowadays to quality assessments stems to a considerable extent from the role that these evaluations can play in defining research strategy.

Therefore, rather than carrying out evaluations of research institutes and groups on its own, the Dutch Government has confined itself to triggering evaluations, to be performed, or at least commissioned by, the research organisations themselves. The typical model for such evaluations is one of external evaluations mostly undertaken by independent peers, often preceded by self or internal evaluations. Sometimes peer reviews are backed up by bibliometric analyses.

In contrast with the institutional evaluations, that have been primarily conceived as a tool for formulating research strategy and not primarily as an instrument for (re-)allocating funds, evaluations of research programmes have been used for financial decision-making on the launching of research programmes (or starting research projects). In fact, many evaluation activities in this field have arisen not only from the need to justify the allocation of (public) funds for launching programmes but also for continuing or ending them. This is because government or NWO-sponsored programmes are often envisaged as “seed funding”; in other words: to create and strengthen nuclei for scientific research in the expectation that successful research, over a time span of four to eight years, will have gained enough momentum to continue on its own strength; and in the expectation that funding will be taken over by other parties than the initial sponsors. Initially programmes seem to have been evaluated “in house” and in a rather informal way by the responsible agency. Later on external and more formal review became the vogue.

From the mid-seventies onwards, more and more attention was paid to assessment of research in specific sectors and to the relevance of research from a societal perspective.

The trend towards a more profound orientation on the (potential) societal relevance of (strategic) research called for formulating (societal) research priorities. This in turn asked for new types of research and research assessments as well as new fora to discuss the societal research priorities. Thus new actors and bodies as well as research evaluation and assessment practices could be identified.

With regard to the assessment of research in specific sectors the first to be mentioned are the “Fact finding committees” that were established in the mid-seventies as a science policy instrument. These committees were engaged in evaluation, rather than in research foresight.

In the beginning they focused on research in areas that were thought important from a societal perspective and evaluation also focused on societal relevance. From about 1980 their focus shifted from these areas to disciplines, and from societal relevance to scientific relevance.

During the mid-seventies, another instrument of science policy was also created: the Sector Councils. These councils, established for a small number of areas, are still in existence and function on the basis of tripartite discussion between those involved in research, the users of research and policy makers.² They are to “steer” the research in areas of societal importance. To underpin this steering, the councils could also initiate evaluations. From 1984 they have become overarched by the Consultative Committee of Sector Councils (COS) that is their forum for discussing issues of mutual interest.

Because of the need for a continuous process of research foresight exercises and the definition of research priorities on the national level, the minister of Education and Science in 1992 established the Foresight Steering Committee (OCV) (Ministry of Education and Science, 1990).

This independent committee was asked to assist the minister in the formulation and implementation of strategic choices and to develop a systematic long-term focus for priority setting in scientific research, based on developments in S&T and the needs of society.

Evaluation of research institutes and research groups

Universities

According to the Netherlands Higher Education and Research Act (WHW), Dutch universities are held responsible for initiating and monitoring their own systems for quality control. The Act contains a number of provisions to ensure quality control. These apply to both teaching and research institutions. Apart from regulations concerning quality control the Act also offers the opportunity to attach conditions to the funding of research at universities in relation to quality assurance (Ministry of Education and Science, 1993).

As a result of the agreements made in 1992 by the Minister of Education and Science and the 13 universities (9 general, 3 technical and 1 agricultural university), the Association of Universities in the Netherlands (VSNU) was asked to set up a system for external research evaluation as a complement to the internal efforts with regard to quality control. The resulting system, partly based on the system for evaluation of educational programmes, is intended primarily to be used as an instrument for research management by the universities themselves, and seeks to underpin decision-making at the various levels

2. There are Sector Councils for *Agricultural Research* (NRLO); *Research in Development Problems* (RAWOO); *Health Research* (RGO); *Research on Nature and Environment* (RMNO); R&D network for *Spatial Planning Policy* (NRO).

within the universities up to the level of the executive board. It involves a review by (predominantly foreign) peers, sometimes based on bibliometric research. As shown in Box 1, the system covers all academic research, is disciplinary oriented, and addresses research at the programme level.

Box 1. VSNU-Evaluations Technical set-up ³
<ul style="list-style-type: none"> • <i>Scientific quality</i>: quality of dissertations & publications; originality and coherence of research; contribution to the development of the discipline; (inter)national position of research; size of research funding acquired from NWO. • <i>Scientific productivity</i>: research output related to human and material resource input. • <i>Relevance</i>: significance for development of the discipline (scientific relevance); or significance with regard to societal/technological impact (societal relevance). • <i>Long term viability</i>: relates to the direction research is taking and to competitive strength that may depend on factors of scale and the scientific infrastructure available. • On each criterion it applies a five-point scale, mostly labelled as: poor/unsatisfactory/satisfactory/good/excellent. • Site visits are for the beta and technical sciences; for other disciplines there are discussions with university representatives at a central location. • Formal appeals are provided for - and in fact have been put forward in relatively few cases - although there is no independent last resort appeal body. • Evaluation reports are publicly available (all in English). • VSNU also publishes a yearly report on the evaluation process for both research and educational evaluations (overviews evaluations carried out and deals with financial aspects and publicity).

Source: Author.

In contrast with the VSNU-evaluations for education - for which there is an Education Inspectorate to monitor the quality of the quality control systems set up individually and jointly by the institutions (meta evaluation) - there is no such monitoring mechanism for research evaluations.⁴ This is why VSNU in 1996 initiated a study to monitor the follow up of the disciplinary assessments. The outcomes of the study - performed by the CHEPS - suggest that this follow up is satisfactory and showed it to be multifaceted, entailing active use (basis for policy choices) and passive use (reports being read and discussed) (Westerheijden, 1996). According to the report none of the evaluation reports have been ignored and all faculties had drawn up action plans to make improvements. Active use seems to prevail. Negative sanctions could also be identified. As yet the follow up did not - at least in most cases - result in reallocations, thus reflecting the previously mentioned characteristic that organisational research assessments are not (primarily) meant as a basis for allocative decision-making.

NWO and KNAW

The Act on the Netherlands Organisation for Scientific Research (NWO) reflects the choice for decentralised decision making. In 1988, NWO replaced the Netherlands Organisation for Pure Scientific

3. It was decided to have a 'trial exercise' for four disciplines in 1993. In 1994 on the basis of these experiences the evaluation protocol of 1993 was somewhat revised.

4. The Education Inspectorate may also carry out evaluation activities of its own in order to assess quality directly, and may appoint a committee of independent experts for this purpose.

Research (ZWO), which dated back to 1950. The small change in the organisation's name was accompanied by major changes in its objectives, ambitions and procedures, as well as in its organisational structure. While continuing to focus on universities and para-university institutes, NWO's sphere of activity was expanded to include the promotion of both pure and applied research in all fields. NWO was also expected to pursue a pro-active stance rather than a passive response to unsolicited research proposals, and to encourage research on the basis of societal considerations. In addition, knowledge transfer was added to its list of statutory duties.

NWO is responsible for promoting quality and innovation in fundamental and strategic research in the Netherlands. Its activities include the funding of university research programmes, projects, research trainees and research equipment, and the provision of scholarships. NWO is also responsible for the exploitation of a number of fundamental research institutes, in physics, astronomy, mathematics, marine sciences, space science and history.

The Governing Board of NWO allocates funds to six Area Councils, which in turn allocate resources to the various NWO-Foundations, which cover the entire range of disciplines, comprising for instance the NWO Technology Foundation (STW) which seeks to stimulate eminent technological research and the application of its results, the Foundation for Fundamental Research on Matter (FOM) which dominates Dutch research in the field of physics, and the Foundation for Historical Sciences. In 1997 NWO received 515 million guilders from the Ministry of Education, Culture and Science. Of its total budget (614 million in 1997) approximately 60 per cent is directly spent on university research projects. A substantial amount of NWO funds is also allocated to outstanding scientists, for instance through programmes like SPINOZA and PIONIER. PIONIER enables young, talented researchers (up to 40 years) – mostly based at universities – to set up or expand research of great (inter)national importance with a team. SPINOZA is designed for researchers (up to 55 years of age) who are internationally recognised leaders in their fields.

The NWO-Act prescribes a periodical evaluation process for the organisation as a whole and its constituent parts. To be more precise the NWO-Act demands that the General Board report once every six years on the achievements of NWO goals. Furthermore, the act also provides for predominantly financially oriented monitoring by the General Accounting Office. The NWO Act also refers to a by-law. In this by-law, which is to be defined by the General Board of NWO but needs approval by the Minister of Education, Culture and Science, the General Board of NWO laid down "manuals" for the set up of evaluation activities in order to stimulate a uniform and systematic approach to these activities. The specifications of the manuals should be considered as minimum-requirements (see Box 2).

In 1995 the Minister of Education, Culture and Science initiated an evaluation of NWO's performance during the first seven years of its existence. An international evaluation committee was appointed and provided with a list of questions (terms of reference) addressing various aspects of NWO's activities. The committee made use of a self-evaluation report of NWO as a major input for its assessment. The committee concluded that NWO is widely respected for the way it executes its primary tasks. Nonetheless, the committee made 12 recommendations, e.g. to increase the NWO budget (for which shifting of part of the university budgets to NWO was considered an option); to create a more active role for the minister of Education, Culture and Science in defining societal priorities for research; to create a two-level organisation (the Foundations level should be abolished); to create an explicit NWO institute policy by bringing the NWO-institutes directly under the authority of the Governing Board and by

designing a single evaluation procedure for both the Academy (see KNAW) and the NWO institutes; and to introduce *ex post facto* evaluation of research projects.⁵

Like NWO, the Royal Netherlands Academy of Arts and Sciences (KNAW) is in practice also subjected to periodical evaluation (see Box 2). There is no separate act on the KNAW, as this body is covered by the WHW. Apart from its task as an advisory body to the government on basic research, it runs its own institutes and provides scientists with facilities to establish national and international contacts and functions as a “meeting” place for scientists. In addition, the KNAW funds post-graduate research at universities. The KNAW budget amounts to Gld 122 million, of which circa 60 per cent is allocated to its own institutes. As the KNAW institutes focus on the humanities, social sciences and life sciences, there is little overlap between research fields covered by these institutes and those covered by the institutes of NWO.

Box 2. Evaluation NWO & KNAW Institutes Technical set-up
<p>NWO</p> <ul style="list-style-type: none"> • Peer review, sometimes backed up by bibliometric analyses. • Evaluation set-up may differ for organisational levels and vary according to institute’s mission and its life-cycle phase; the following characterisation applies as a rule. • The scope covers aspects like: vision; strategy; scientific quality; R&D; HRM and financial management; quality and coherence of the institute programme; infrastructure and housing; position of the institute; size and quality of so-called third-party funding; follow up of previous recommendations. • Site visits are standard procedure. • Review committees: consist of at least three peers; the chairman or at least one of the peers is Dutch. • Evaluations start off by self-assessment and may address past performance as well as the expected research potential. • Follow up is embedded in managerial processes and discussions on the level of the executive board of NWO. Reports may end up in allocative shifts. • Evaluations are carried out five-yearly. • Sometimes ad hoc evaluations are initiated. • Almost every institute is counselled by an external committee. <p>KNAW</p> <p>As a rule the same applies to the evaluation of the KNAW institutes, the exceptions being that KNAW:</p> <ul style="list-style-type: none"> • Focuses more on scientific quality than on societal relevance related criteria. • Uses peer review committees, chaired by a Dutchman, which are expected to use a fixed-interval (4 point) assessment scale. • Sends summaries of its evaluation reports to the minister (for information purposes only).

Source: Author.

5. In the AProtocol 1994, *Quality Assessment of Research* (VSNU, Utrecht, 1994), a complete description is given of the responsibilities and tasks of the different parties involved in the assessment process.

At the beginning of 1997 an evaluation was initiated by the Minister of Education, Culture and Science with regard to the organisational structure of KNAW and the functioning of its bureau. This evaluation will be carried out by an external evaluation panel of two foreign experts, chaired by a Dutchman, preceded by a self evaluation.

Research schools

Apart from these evaluation activities KNAW is also involved in another kind of institutional evaluation: the accreditation of so-called research schools. PhD-students under the guidance of university staff carry out a sizeable portion of university research. In order to offer better training programmes to the PhD-students and to create stimulating research environments for research training, these research schools have been introduced within the university system since 1990. Proposals for setting up research schools are to be submitted by the universities, which continue to develop their own research “profile”. Through the research schools co-operation within universities, between universities and with other institutes has increased significantly.

Research schools must be accredited by a special independent accreditation committee (ECOS) which has been brought under the auspices of the Royal Netherlands Academy of Arts and Sciences (KNAW) because of 1) its advisory role mentioned earlier; 2) the notion that the monitoring of effectiveness should be separated from the allocation of funds; and 3) because NWO itself at the time was involved in the allocation of funds to research schools. To receive accreditation, research schools must meet various criteria in terms of quality, organisation and scope (see Box 3). As yet 97 research schools have been accredited by the ECOS.

Box 3. ECOS-Accreditation of research schools Technical set-up
<ul style="list-style-type: none"> • ECOS assesses primarily on ten criteria relating both to the content of accreditation proposal and some additional aspects. In a nutshell these criteria relate to the: <ul style="list-style-type: none"> - Coherence of the (educational and) research programme; - Organisational set up (from a managerial perspective); - Financial guarantees/backing; - Mission definition; - Size (minimum requirements). • Accreditation will be evaluated after five years on the basis of: <ul style="list-style-type: none"> - Firstly the same criteria as were applied by the first accreditation; - Secondly criteria regarding the functioning of the research school since the first accreditation: quality and productivity; policy implementation; motivation of policy shifts; follow-up of previous recommendations. • ECOS consists of nine persons and is assisted by seven subcommittees covering all scientific disciplines; protocols for each of these sub-committees vary. Subcommittee members are authorities within the discipline-community. • ECOS reports in writing. Universities may be asked to elucidate proposals orally.

Source: Author.

TNO

In the Act on the Netherlands Organisation for Applied Scientific research (TNO), no requirements are specified with regard to research evaluation (although the act provides for the financial monitoring by the General Accounting Office, counselling by the Programme Advisory Councils, and monitoring of management policy by the Board of Governors). The primary task of TNO is to carry out strategic and applied scientific research serving the central and lower authorities, the business sector and other social groupings. TNO runs 16 institutes in the fields of industry, nutrition, health, environment and energy, transport and infrastructure, building research and defence research. The organisation undertakes policy studies as well. Its work force accounts to more than 4 000, its budget is about Gld 750 million (1997).

The TNO budget can be divided into three main parts:

- ◇ Basic funding (about 15 per cent of the TNO budget) from the Ministry of Education, Culture and Science. This is meant as an instrument for research-policy of the TNO Board of Management with which TNO can enter into high risk research of an exploratory nature in an early stage of development and for entering new fields (fostering the knowledge base). A multi-year programme was recently presented to the Minister of Education, Culture and Science.
- ◇ Programme funding (29 per cent of the TNO budget) from the other ministries, used for specific multi-year research programmes for the different ministries. In these programmes TNO's expertise is geared towards fields the ministries consider important in the medium term. Programmes are proposed by TNO and, after negotiating, TNO and the ministries make up yearly bilateral agreements on these programmes.
- ◇ Funds from contract research (56 per cent of the TNO budget) which must be cost-effective and funded entirely by the client (private sector, government, international bodies).

Because the first and second part of the TNO budget are allocated on a revolving basis, they offer some opportunities for introducing evaluative elements in fund allocation to TNO. Furthermore, TNO, according to the TNO act, has to draw up a strategic plan every four years that should reflect assessment of past, present and future performance. In 1996, TNO started to "audit" its institutes. Box 4 specifies the set-up of these audits.

Large technological institutes

From the twenties onwards several large technological institutes were set up because of the need for specific expertise in areas of research demanding investments in large facilities: the Hydraulic Laboratory (WL, est. 1927); the Maritime Research Institute in the Netherlands (Marin, est. 1929); Delft Geotechnics (GD, est. 1934); the National Aerospace Laboratory (NLR, est. 1937); and the Energy Research Center in the Netherlands (ECN, est. 1955). The mission of these institutes is defined by the various ministries responsible for them. Every four years these institutes and their mission are evaluated. Their mission statement is reviewed on the basis of the evaluation results.

It is important to stress that the quality of these institutes, and many other institutes which carry out research of a more applied character, can not be defined in terms of the scientific quality of their research only. The organisation, infrastructure, and the specific threats and opportunities within the setting in which they operate should also be taken into account. This is why these institutes are evaluated on a

flexible, “tailor-made” basis, as reflected by the three evaluation criteria applied: scientific, societal and operational quality (see Box 5).

Box 4. Evaluation of TNO Institutes Technical set-up
<ul style="list-style-type: none"> • Institute evaluation is closely linked to the “Technology Portfolio System”. • Through this system TNO monitors its research on specific themes, relevant to the market, in order to foster its strategic knowledge base. • This knowledge base is defined in terms of “expertise nuclei” (“We know how to.....”). • Technologies are rated in terms of their stage of development, the strengths as compared to technologies of competitors; and “market attractiveness”. • Each technology is “anchored” in a research group. • Each technology turnover is monitored. • International committees review the institute portfolio on the basis of an external audit, which looks both at quality and market relevance. • Portfolios are also discussed twice a year by the TNO Board and institute management teams. TNO Board uses audit results as an input for defining R&D strategy.⁶

Source: Author.

Box 5. Evaluation of the large technological institutes: Technical set-up
<ul style="list-style-type: none"> • Evaluation practice tends to be in accordance with the set-up as defined in the Science Budget of 1991. • Three evaluation criteria are to be applied: <ul style="list-style-type: none"> - <i>Scientific quality</i>: the quality and originality of scientific and technological research. - <i>Societal quality</i>: mission of the institute; applicability of research findings in practice. - <i>Operating quality</i>: abilities on the level of implementation and process management; efficiency and knowledge exploitation. • Parties involved: party responsible for evaluation, the evaluating institution (often a committee), institute sponsor(s), consultants and a sounding board panel, apart of course from the institute that is evaluated. Sounding board committees define criteria and underpin credibility of the evaluation process. • Evaluations for these kind of institutes should be tailor-made exercises, sometimes building on self-evaluation or on ad hoc evaluations (investment decisions). • Evaluations of the functioning of the institutes as a whole – such as the ones above – should be distinguished from evaluations concerning specific functions. In these specific evaluations “standing scientific committees” play a role. • Institutes and relevant ministries discuss evaluation results on the basis of the strategic plan.

Source: Author.

6. In 1996 audits were carried out for the TNO Road-vehicles Research Institute; TNO Human Factors Research Institute; and in early 1997 for TNO Nutrition and Food Research Institute. In 1997 audits for the TNO Building and Construction Research Institute, the TNO Prins Maurits Laboratory, and the TNO Institute of Environmental Sciences.

Programme and project evaluations

Programme evaluations

Many evaluation activities in this field stem from the need to justify the allocation of funds not only for launching but also for continuing programmes, that often are intended as “seed funding”. Therefore, evaluations may be *ex ante*, *ex post facto* or midterm.

Ex ante evaluation may involve *programme* proposals or proposals for *projects* within these programmes.

The objective of “seed funding” perhaps particularly applies for the public funded programmes, such as the National Research Programmes that were operational in the seventies, the Innovation Oriented Research Programmes (IOPs), and the Science Policy Spearhead Programmes which were launched from the mid-eighties and well into the nineties.

In the beginning these programmes were often evaluated “in house” and in an informal way by the responsible agency, although in several cases – like in the case of the Information Stimulation Plan (INSP) – evaluation involved external and independent scrutiny.

However, as time went by and also new programmes were launched which focus on research of perceived societal relevance, gradually a more systematic, thorough and formal assessment practice for these programmes emerged, based on the assessment by external review committees, empanelled by representatives of the research and business community, as well as other sectors of society.

For some time past Innovation Oriented Research Programmes, aimed at strengthening the research base in innovation relevant areas, are evaluated on the *programme level* systematically *ex ante*, mid-term and *ex post facto*.

Ex ante evaluation encompasses: the selection of themes; preparation of programme proposals; and go/no go assessment.

Projects within the IOP programmes are evaluated *ex ante* (selection of project proposals) and *ex post facto* (assessment of results), on scientific quality, correspondence with the themes selected and interest of enterprises.

IOPs are evaluated externally (often by a consultancy company) in terms of goal achievement.

In accordance with the principle of “government at arm’s length”, the Science Policy Spearhead Programmes were set up by independent programme committees and were often evaluated by independent bodies like the KNAW or Sector Councils. In 1991 the Spearhead Programmes were also evaluated positively as an instrument of science policy by the General Accounting Office. The Science Policy Spearhead Programmes preceded the current Incentive Programmes.

As a rule these strategic and innovation oriented programmes show a pragmatic mix of expert evaluation backed up by systematic data gathering; sometimes a preliminary study – intended as a zero-base assessment – is launched (Rip and Van der Meulen, 1995). Evaluation for these strategic programmes encompasses both an assessment of scientific quality and an “impact evaluation” with regard to the quality of research from a societal perspective.

Apart from the strategic programmes, like the PRIORITEIT Programmes that encourage fundamental strategic research in fields of perceived societal relevance, NWO runs several other programmes that focus on the HRM-aspect, such as the PIONIER and SPINOZA programmes (see Box 6).

Box 6. Evaluation of programmes by NWO: Technical set-up
<ul style="list-style-type: none">• Although the methodology may differ, a protocol is provided for.• As a rule evaluation is carried out on the basis of mid-term and end-reports, often backed up by self-evaluations, which are submitted to an external committee that partly consists of foreign members.• Evaluation reports are submitted to the Councils and General Board.• For the PIONIER and PRIORITEIT programmes, evaluation mechanisms have been laid down.• For PIONIER yearly and biennial reporting and evaluations are carried out mid-term and <i>ex post facto</i> by the council and foundation boards. Sometimes a separate evaluation committee pays a site visit and as a rule external advisors are also involved.• For PRIORITEIT <i>ex post facto</i> evaluation preceded by a mid-term evaluation after three years is the rule.

Source: Author.

In 1986 the KNAW started the Academy Researchers Programme which aims at ensuring that potential top-researchers will not prematurely step outside the arena of academic research (1995 budget: Gld 18.6 million.). This is the only major programme run by KNAW. In 1993 the General Board of KNAW instigated an external evaluation to monitor the effectiveness of the programme in terms of its objectives. The evaluation report, that was based on many interviews, concluded that the programme had functioned well, and that the set up only needed some small adjustments. The General Board of KNAW decided to continue the programme and discussed the report with VSNU and NWO.

Project evaluations

Evaluation of project proposals

For project evaluations it is even more difficult to define common characteristics than for the other types of evaluation. Project proposals are evaluated on the basis of jury-reports (in writing) and/or peer review committee discussions and/or judgements by the Programme Committee. Project evaluations are often accompanied by consultations of external, preferably foreign referees, and experts from outside the research community. The jury often comprises a “laymen panel of wise men” (wise men who are not necessarily experts in the relevant field); the referees encompass mostly experts in the field and users of the research concerned.

In 1993 the executive board of NWO specified a general protocol, the details of which are depicted in Box 7. Several NWO Foundations, such as FOM, have a somewhat different set up for these evaluations.

Box 7. Evaluation of *project* proposals by NWO

Technical set-up

- Firstly for each proposal at least two (preferably foreign) referees should be consulted.
- A “jury” (reports only in writing) or a committee (reports on the basis of discussions) evaluates the proposals on the basis of the referee-assessments and – if necessary – defines priorities; the jury composition covers the tasks and working area of the NWO Foundation.
- Jury/committee-members will be appointed on a yearly basis; 2/3 at the most may be re-appointed directly; members may be re-appointed three times at the most.
- When useful, external experts from outside the academic community may be consulted.
- The Foundation Board or Area Council Board decides on proposals.
- Procedures may not exceed five months.

Source: Author.

As pointed out, it is the responsibility of NWO-STW to stimulate eminent technological research and the application of its results. Although each of the various NWO-Foundations are involved in project assessment, NWO-STW is interesting because of its specific approach: apart from selecting projects on the basis of scientific quality, STW also evaluates projects with regard to “utilisation”. The latter refers to the chance that users of the research outcomes will apply these results. The utility of these applications plays no role in the selection process (see Box 8).

Ex post facto evaluations

According to the report of the NWO evaluation committee, “NWO does not have a procedure for *ex post facto* evaluation of research projects. Of course, whenever a scientist applies for funding, the results of any previous work funded by NWO are considered. However, this practice does not produce systematic information about the quality of the assessment procedures followed by NWO” (NWO, 1996). Evaluation of NWO projects is often done on the basis of final reports.

STW not only evaluates projects with regard to the “utilisation” on an *ex ante* basis but also five and ten years after they have started. The purpose of these evaluations is to account for the allocation of funds, to “create legitimacy” and to improve insight into the “utilisation”. It is important to stress that STW developed an approach which seeks to improve the “utilisation”. This is exemplified by the Utilisation Committees for each of the projects that gather every half year for monitoring and improving utilisation. Two years after commencement, the General Board of STW officially decides on the continuation of the project. This decision is based on discussions with the Utilisation Committees. STW publishes the results of these monitoring activities and the follow up of projects in its Utilisation Reports.

Apart from this, STW also brings together research projects that share a common application area to improve the “visibility” of these projects and to stimulate mutual fructification.

Box 8. Evaluation of project proposals by STW

Technical set-up

Proposals may be submitted continuously; evaluation is started when 20 proposals are available and involves two rounds.

FIRST ROUND

- First there is an anonymous *peer review* by five to eight experts in the relevant field and from sectors of potential users.
- Applicants may react to these reviews. The results are added to the proposals.
- The scientific and “utilisation” criteria to be applied are specified in a questionnaire.
- Scientific criteria covers aspects such as the:
 - expertise of the team;
 - originality of the proposal;
 - effectiveness of the method.
- Utilisation relates to matters like:
 - applicability in sectors of society, like industry;
 - opportunities for Dutch companies to produce commercial products;
 - competitiveness of Dutch Industry;
 - importance for technological progress;
 - position of Dutch industry with regard to patents.

SECOND ROUND

- This round involves a Delphi procedure (in writing) by which the outcomes of the first round (20 proposals, peer review, comments) are submitted to 10-12 jury members: a “laymen panel of wise men” (wise men who are not necessarily experts in the relevant field).
- Jury members should apply both scientific and “utilisation” criteria.
- They give one score for each of these aspects.
- The results are submitted to jury members who then may adjust their assessment.
- Scores are then averaged. The average scientific quality and utilisation quality are then averaged in one ultimate assessment.

Finally the STW Board decides to fund the eight projects with the highest score.

Source: Author.

Research evaluation, societal appraisal of research, and research foresight

Since 1932, the year that Huxley’s *Brave New World* was published, and 1933, the year of the World Fair in Chicago, that had as its slogan: “Science explores, Technology executes, Man conforms”, the relationship between science and society has become more complicated (Foresight Steering Committee, 1996). During the second World War, science was used as an important strategic resource for government policy and society at large. The very positive promise, “Science, the endless frontier”, in Vannevar Bush’s words, led to sustained growth in government funds for science, while at the same time research steering was mainly left to the scientists and their organisations. The last decades however have shown a marked development towards societal steering of research: science and technology are nowadays seen as very instrumental for creating wealth and well-being in society, so that science and government technology policies are now oriented more to setting priorities for research and to embedding research in its societal environment. The answer to the question “what is important for society?” has also become more decisive for the societal appraisal of science and technology.

The sectoral level

The trend towards a more profound orientation on the (potential) societal relevance of research asked for formulating societal research priorities and coincided with the emergence of research foresight exercises. This in turn asked for new types of research and research assessment as well as new fora to develop and discuss possible societal research priorities, e.g. through foresight processes. Thus, new actors, bodies and research evaluation practices could be identified.

This is where the Sector Councils come in, which have been in existence for about 20 years now and function on the basis of tripartite discussions between those involved in research, the users of research and policy-makers. Once every four years they produce a Long-Term Perspective, based on a thorough investigation of relevant social and scientific trends. Sector Councils are financed by the minister(s) responsible for the sector concerned.

At present there are Sector Councils for *Agricultural Research* (NRLO); *Research in Development Problems* (RAWOO); *Health Research* (RGO); *Research on Nature and Environment* (RMNO); R&D network for *Spatial Planning Policy* (NRO).⁷

The task of the Sector Councils is to develop a multi-year view on sectoral research policy and priorities based on scientific and societal issues and on an optimal match between supply and demand in the different fields of research for which they are responsible. An important activity of the Councils in this respect concerns the translation of societal issues into a well defined programming of scientific research.

Although evaluation is not a principal task of the Councils, four types of activities in the context of assessment can be distinguished. The Councils:

- ◇ stimulate discussions on the assessment of research from a societal perspective (criteria and methods to be applied);
- ◇ advise on criteria and organisation of evaluation procedures in a sector;
- ◇ occasionally play a (formal) role in regular assessment processes (assessment from societal perspectives only);
- ◇ analyse the scientific quality and societal relevance of research in their sector in terms of strengths and weaknesses, when formulating their multi-year view.

Such bodies and their activities, involving research foresight exercises, constitute another typical characteristic of the Dutch model, as they capitalise on – and draw heavily from – consensus building and mutual accommodation; a way of decision-making that is also typical for many discussions in other fields. The disadvantage of this bottom-up approach, of course, is that policy formulation may be time consuming and easily leads to “middle of the road” conclusions. The obvious advantage however is that policy implementation can be more effective as well as less time consuming as all parties support the choices made.

7. Although it does not meet precisely the requirements of the Sector Councils Framework Act, adopted in 1987, the NRO also participate in the system, because of the close commonalities in responsibilities.

The national level

The Dutch Government became interested in foresight exercises after the publication of Irvine and Martin's "Foresight in Science" in 1984 (Rip and Van der Meulen, 1995). It commissioned a study by the same authors in 1989 (Irvine and Martin, 1995).

Because of the need for a continuous process of research foresight exercises and the definition of research priorities on the national level, the minister of Education and Science in 1992 established the Foresight Steering Committee (OCV) (Ministry of Education and Science, 1990). This independent Committee was asked to assist him in the formulation of strategic choices and to develop a systematic long-term focus for priority setting in scientific research, based on trends in S&T and the future needs of society. The OCV was asked to present options for both priorities and posteriorities, and to develop a consensus among universities and research organisations to promote implementation of these options in the policy making of these institutions. The activities of the OCV focus on the whole spectrum of publicly funded research. To produce options for the total volume of publicly funded research, the OCV has used exploratory surveys and foresight activities (studies, workshops, conferences, etc.) along disciplinary lines and in areas of specific interest for societal sectors/issues (see Box 9).

Box 9. Research appraisal by OCV Technical set up
<ul style="list-style-type: none">• The <i>Foresight Steering Committee</i> performed its tasks by:<ul style="list-style-type: none">- carrying out foresight studies itself, or commissioning such studies;- workshops, conferences, etc.;- consulting relevant bodies and experts from science and society;- commissioning other foresight activities;- regular discussions with the minister.• The Committee adopted the use of a framework (OCV-matrix) that gives a stratified picture of the present and the future research at the educational, the sectoral and the societal level, the options that exist, and the activities that will need to be undertaken.• This framework relates to:<ul style="list-style-type: none">- elements of societal demand, i.e. the relationship between research and education and training,- sectoral interests and cross-sectoral societal issues;- elements of the foresight process, i.e. developments in a research field.• Foresight studies also take international orientation into account, as well as contacts and collaboration with other fields of research.

Source: Author.

The final OCV report – published (also in English) in 1996 – constituted a major input for the Science Budget 97, the central policy document in Dutch science policy (Foresight Steering Committee, 1996; Ministry of Education, 1996). For each field of knowledge and area of science, the Cabinet indicates which concrete choices it has made and which programmes will receive more funding. It is important to stress that the OCV report not only sought to underpin the Cabinet's Science policy but also – and most importantly – was intended for the various players in the research arena to take the OCV recommendations into account when formulating and implementing their own policies. In the context of the bilateral discussions with, for example, universities, this follow up will be discussed and monitored.

The final report is based on a large number of "foresight" reports by expert committees, which the OCV has had carried out in a variety of scientific fields in the past few years.

Recently the Ministry of Education, Culture and Science launched a study in order to take stock of (best) practices for evaluating basic and strategic research from the perspective of societal relevance. The results of this study will be available in April/May 1997.⁸

Studies in evaluation methods have been carried out mainly under the auspices of COS. The Ministry of Education, Culture and Science actively supported the Centre for Science and Technology Studies of the University of Leiden (CWTS) in the development of bibliometric methods. In 1992, the Netherlands Observatory of Science and Technology (NOWT) was established by the Minister of Education and Science to assess the current situation and keep track of trends in science and technology, scientific literacy and attitudes towards science and technology. The prime objective of the (biennial) NOWT indicators reports is to compare the Dutch scientific performance with the international situation on the basis of quantitative indicators.⁹

Looking ahead

Some remarks are to be made about the ways in which evaluation actors, the Minister of Education, Culture and Science included, try to “harmonise the winds of change”.

Bridging evaluation practices

The enrichment of Dutch research evaluation culture mentioned in the introduction raised questions regarding the possibilities for, and necessity of, improving coherence of the assessment practices and countering undue overlap. The CHEPS study on the follow up of the disciplinary evaluations, which was carried out for the VSNU, corroborated the growing concern for assessment fatigue, owing to the substantial workload as a result of the numerous evaluation efforts: VSNU-visitations; ECOS-accreditations; NWO/KNAW programme and institute evaluations; foresight exercises; and inventories or evaluations, etc. for various fields of research, e.g. energy research, by a number of actors.

In fact, this concern was anticipated. In the Higher Education and Research Plan (“HOOP” in Dutch) of 1996, the Minister of Education, Culture and Science announced that he would ask VSNU, KNAW and NWO to advise him on the issue of bridging their evaluation practices (Ministry of Education, Culture and Science, 1995). At the close of 1996, a tripartite working party of these organisations published its report (VSNU, 1996). In a nutshell the primary recommendations to NWO, KNAW and VSNU of the working party are:

- ◇ opt for a six yearly (instead of five yearly) cycle for the VSNU evaluations and set an evaluation agenda for each cycle;
- ◇ improve the alignment of the set up of research and educational evaluations;
- ◇ use more common definitions when asking for information that is used as an input for the various evaluations;

8. The study is carried out by Rip and Van der Meulen, University of Twente.

9. Netherlands Observatory of Science and Technology (NOWT). *Wetenschaps- en technologie-Indicatoren 1996. English summary: Science and Technology Indicators 1996*. NOWT is a joint project of: the Centre for Science & Technology Studies (CWTS, University of Leiden); The Maastricht Economic Research Institute on Innovation and Technology (MERIT, University of Maastricht); and the Dutch Agency for Research Information (NBOI).

- ◇ avoid as much as possible additional workload for researchers as a result of the (re-)accreditation of research schools and allocate tasks regarding the accreditation processes as efficiently and effectively as possible.

It seems sensible, when defining the overall set-up of the evaluation system for the years to come, to also consider some potential “deficiencies” in – or consequences of – the set-up of current evaluation practices, both on the level of process and content. With regard to the VSNU evaluation system, it seems appropriate to face, for instance, the question of in which way the mono-disciplinary VSNU system of research evaluation should be balanced with the need for evaluating multi-disciplinary research that often focuses on societal needs.

Towards an output oriented funding

As yet, research evaluation is not linked to government funding of university research. The funding of the universities is based on the number of students on the one hand, and on the research of the universities on the other hand. This latter “research component” consists of several compartments:

- ◇ a basic provision component (not earmarked);
- ◇ a dissertation/designer-certificate component;
- ◇ a research school component;
- ◇ a “strategic policies” component.

In the Higher Education and Research Plan of 1996 it was announced that the dependence on student influx numbers was to be decreased and the funding of university research was to be stabilised to a certain extent (Ministry of Education, 1995). In this situation it is of course very important to ensure a high quality level of research as much as possible.

If output-oriented funding of university research has to be introduced, then scientific quality and – to some extent – criteria related to the orientation toward societal goals may be used for funding university research. Research evaluation then will have a key function in the funding process, as is the case in, for example, the United Kingdom.

REFERENCES

- FORESIGHT STEERING COMMITTEE (1996), *A Vital Knowledge System; Dutch research with a view to the future*, Amsterdam.
- IRVINE and MARTIN (1989), *Research Foresight; Priority Setting in Science*, Pinter Publishers, London.
- MINISTRY OF EDUCATION AND SCIENCE (1990), *Science Budget 1991*, p. 13, Zoetermeer.
- MINISTRY OF EDUCATION AND SCIENCE (1993), *Everything you always wanted to know about the Higher Education and Research Act (but were afraid to ask)*, p. 19, Zoetermeer.
- MINISTRY OF EDUCATION, CULTURE AND SCIENCE (1995), *Ontwerp Hoger Onderwijs en Onderzoekplan 1996*, Zoetermeer.
- MINISTRY OF EDUCATION, CULTURE AND SCIENCE (1996), *Science Budget 1997*, Zoetermeer.
- NWO (1996), Report of the NWO evaluation committee, p. 28, The Hague.
- RIP, A. and B.J.R. VAN DER MEULEN (1995), "The patchwork of the Dutch evaluation system", *Research Evaluation*, April.
- VSNU (1996), *Afstemming onderzoekbeoordelingen*, Rapportage Tripartite werkgroep afstemming onderzoekbeoordelingen, Utrecht, (only available in Dutch).
- WESTERHEIJDEN, D.F. (1996), *Een solide basis voor beslissingen*, Effecten van de onderzoekvisitaties door de VSNU, Center for Higher Education Policy Studies (CHEPS), University of Twente, Enschede.

CHAPTER 3. EVALUATION OF SCIENTIFIC RESEARCH IN THE UNITED KINGDOM

*Derek Barker and Philippa Lloyd, Head of Research Management and Manpower Division
Office of Science and Technology, United Kingdom*

Introduction

Basic research provides the basis for future applied research and is particularly important in training scientific researchers and enabling teams of scientists to attain and stay at the leading edge in their areas of study. Without basic research there is no applied research; without applied research, there is no technology and without technology there is no success in the marketplace. As such, basic research provides a key basis for the competitiveness of individual firms, academic institutions, for regions, and countries.

Given the importance of basic research, it is not surprising that there is great interest in developing ways for evaluating the outcomes. Such mechanisms, if sound, can helpfully guide policy decisions on the future support of basic research. However, precisely because of the inherent nature of basic research, which is generally concerned with gaining insights into understanding and investigating phenomena, rather than with developing specific outcomes such as new products, processes or services, this evaluation is not straightforward.

The UK science base

First, some facts about the UK research scene. Almost all of the basic research carried out in the United Kingdom takes place in the science and engineering base which consists of the UK universities and research institutes. Funding is primarily from the Government, although the charities, especially those operating in the medical field, also contribute significant sums. There are two main sources of government funding for the science and engineering base: there is the block grant which is provided to support infrastructure and the basic capability in universities to carry out research and teaching, and the “science budget” which mainly pays for individual research projects and postgraduate training.

The block grant is administered by the Funding Councils (one each for England, Scotland and Wales, and an equivalent arrangement for Northern Ireland), and each university’s research funding is chiefly based on quinquennial assessments of the international research standing of individual departments in the university. The science budget is administered by the Research Councils, of which there are six concerned with this funding role. In deciding which research projects to fund, research applications are subject to review. Peer review is the system by which the intellectual quality of a grant application is assessed by researchers working in, or close to, the field in question. What is sometimes called “merit review” adds consideration of relevance and value for money to the consideration of quality. Assessments by peer and merit review can be used to decide which projects to support, in assessing the progress of research, and in evaluating the outcomes of projects through, for example, consideration of the required end-of-grant report.

In decision making on “blue-skies” research (termed responsive-mode funding), peer review has the larger role. In modes of support that are explicitly in pursuit of Councils’ strategic aims, merit review criteria comes more to the fore, although consideration of scientific excellence remains the most significant factor. Poor quality research is not merely a poor buy – it can be extremely dangerous. Bad research can lead to bad policies, bad products, bad services. And the financial and other consequences of poor research advice can be disastrous. It is crucial that government-supported research be of the highest quality.

The concept of relevance can create some difficulty; some researchers see relevance as synonymous with low quality. And some see it as a generally undesirable characteristic anyway. But this is to misunderstand the importance of relevance. As the 1993 White Paper on science said, it is essential that excellence in UK science be matched with excellence in turning the outcomes of this effort to useful account. That can mean enhancing the wealth-creation capability of firms, or it can mean enhancing the quality of life of the nation.

Thus, in addition to sustaining the excellence of the science base, we must ensure that we carry out research in areas in which it matters, and we have to ensure that the science base is well connected to those who make use of its outcomes. We must also have a business community able and willing to take the outcomes from the science base, combine these with the fruits of their own research, development, and commercial activities, and get profitable products, processes and services to market.

Assessing relevance is therefore a complete topic in itself, which will not be discussed in detail in this paper. It suffices to say that one of the most well known mechanisms for assessing relevance is the Foresight process. Foresight is not about picking winners. But Foresight, through the process of connecting the science base and users together in a productive manner, can help establish a consensus. Foresight aims to help all those involved reach agreement on at least some important areas in which, taking a 20-year view, there are likely to be major technical developments and significant commercial opportunities. Above all, Foresight emphasizes that underpinning science should be supported.

Reaching consensus is very important, and it can help guide decisions on research priorities, for instance. But it is only one of the factors taken into account by the Research Councils in reaching decisions on which areas of research are “relevant” and merit particular support. Research Councils also cultivate their own networks among the users and providers of research in order to facilitate priority setting.

Value for money is also important in research; we cannot afford to support all the high-quality and highly relevant research within the context of a single nation. But we must strive to ensure value for money so that we can support the maximum amount of research with the available funds. In the United Kingdom, we have pursued a number of initiatives in recent years to ensure that our science base is efficient. Efficiency does not mean simply low-cost – it means that the required goals are achieved with the minimum expenditure of resources. Therefore, we try to ensure that cost considerations are balanced against effectiveness and ability to achieve goals.

Evaluation and its difficulties

There are a number of particular, albeit well-known, difficulties in evaluating basic research. Basic research is far from market and generally divorced from easily characterised or tangible outcomes. It covers a vast range of intellectual endeavours and the boundaries between individual areas are fuzzy or non-existent. Furthermore, the connections between individual pieces of basic research, and between basic research itself and applied research are highly complex. The outputs of basic research can be used in

a number of complex, and often unforeseeable, ways, and the time scales to which the outputs are applied, and therefore their value recognised, are long and can be decades or more.

Despite these difficulties, it is greatly desired that sound evaluation mechanisms be made available.

Evaluation of basic research should help us to address the following issues:

- ◇ Are we funding the right amount of basic research?
- ◇ Are we funding the right basic research?
- ◇ Are we getting optimum value from the basic research we are funding, assuming the research is in the “correct” areas?

Evaluation options

Evaluation of a science base can in principle take a number of forms. The role of the science base is to produce highly skilled and trained men and women and to conduct research at the frontiers of knowledge; this definition of the science base’s function is endorsed by much of industry, in the United Kingdom at any rate, particularly those industries that derive success from their interactions with the science base. Some directly measurable outputs of the science base are therefore knowledge and trained people.

Assessment of the quality of research can be subjective or objective. A common example of the former is to assemble a group of peers of the researchers under review and to seek their collective views about the quality of the research output. An example of the objective approach is evaluation of the quality of scientific papers produced by the researchers by examination of citation data.

Both of these methods are routinely used in the United Kingdom, and both have advantages and disadvantages.

Peer assessment

Peer assessment, which is used for both academic science and for some industrial research of a fundamental nature, is widely used. It has considerable advantages in exposing the ideas and proposals (and, in some circumstances, the achievements, i.e. it can be used for *post facto* evaluation of outcomes) to the scrutiny of others who are skilled in the scientific or engineering area under consideration. It identifies ways in which proposals can be improved, as well as determines which should be supported and which should be rejected. It is not easy to come up with an alternative mechanism for evaluation of research proposals which commands the broad support enjoyed by peer review.

Yet there are those who argue that peer review suffers disadvantages. For instance, it can be susceptible to the prejudices and priorities of the individual assessors chosen. It may also tend to lend more encouragement to research effort that runs along lines of enquiry which are regarded at the time as “sound” or “productive”, thereby tending to act against individuals and groups who pursue research in new areas or who adopt unconventional approaches.

In some areas of science, one can often find two mutually opposed schools of thought - those who favour a particular approach or line of enquiry, and those who do not. This feature means that peer review of this sort runs the risk of, at the extreme, uncritical support for researchers who follow the same approach as that favoured by the assessors, or unthinking dismissal and denial of quality should the reverse be the case.

Despite these potential or actual problems, peer review is the cornerstone of the United Kingdom's approach to determining which research merits funding.

Bibliometrics

The assessment of scientific output by analysis of publications and citations of articles appearing in internationally reputable peer reviewed journals is also widely used. Among its advantages, are that a large international database is available to help perform the comparisons, and there is a considerable amount of expertise available on the relevant analytical techniques. However, some criticisms have been made of bibliometric analysis. For example, it is not regarded as a suitable technique in certain fields, such as economic or social research, where publication and the mechanisms used for dissemination of research findings are not always the same as those used in the natural sciences. It has also been claimed that there may be a tendency for at least some such analyses to work to the disadvantage of those research groups where English is not the language in use. And bibliometric analysis may not give adequate recognition to particularly novel research which may not attract much attention or even support for publication in the early days.

Despite these potential or actual problems, our belief is that bibliometric analysis can provide useful information, provided that care is taken in the analytical approaches adopted. We have made some recent use of this sort of analysis as will be described later.

Assessment of quality of trained people

While such an assessment does not strictly provide an evaluation of basic research, it does shed some light on the quality of the science base. A key output from the science base of vital importance to industry is the supply of trained people. There are, however, few objective measures to gauge their quality. Clearly, gross imbalance in supply, whether overall or within specific disciplines, or gross deficiencies in the quality of graduates will become self-evident. In the United Kingdom, the overall supply and demand for science and engineering postgraduates is, as far as we can determine, roughly in balance. In the medium to long term, of course, restoring forces act to correct gross imbalances, both in the overall supply/demand ratio and in the ratio in individual subjects. Over-supply leads to reduced employment prospects, while under-supply leads to employer demand, both of which influence prospective students' decisions.

Unemployment for science and engineering postgraduates is lower than the UK national average, and there is some evidence that the period out of employment while actively seeking work is also lower.

Determining the quality of postgraduates, and the match between this parameter and employers' requirements is very difficult. It is, for instance, no simple matter to seek the view of past, present and potential future employers on the quality of graduates and postgraduates from individual universities, or in specific disciplines, or even at the level of individual departments. While there is a body of anecdotal evidence, this is of questionable value as it may reflect the prejudices, special experiences or circumstances of individuals, rather than portraying a truly national perspective.

Ideally, nation-wide surveys of the career paths of (post) graduates are desirable to provide reasonably comprehensive and objective data. Experience in the United Kingdom suggests that such exercises are very costly and tend to result in "questionnaire fatigue", which in turn leads to low response rates and dubious quality of response; and such difficulties in obtaining adequate high-quality returns result in data of limited value.

It is widely found that a large number of returns in studies of the initial employment destination of students is “not known”; and many of the returns refer to employment which is unlikely to represent anything more than temporary employment e.g. as supermarket check-out clerks. The second and third career destinations of students may in such cases be much more useful in illuminating the utility and quality of the people and their training, but these facts are much more difficult to determine.

In an attempt to gain more comprehensive information on the utility and relevance of postgraduate training and research, the UK Office of Science and Technology, together with the Research Councils, is funding a study of career paths of students supported in recent years by the Research Councils across all disciplines. This study is due to report in the summer of 1998.

Recent experiences in the United Kingdom

At the national level

At the national level, an evaluation based on a bibliometric analysis of scientific publications and citations has recently been carried out for the majority of the UK science and engineering base. This evaluation, which was aimed at generating a broadly-based comparison with the performance of other countries, was based on the Science Citation database of ISI. The comparisons were carried out for all countries that published 100 or more papers in science, medical or engineering journals indexed by the ISI over the period 1981-94. The data were broken down into 20 fields of science.

The results showed that the United Kingdom produced 671 944 papers, or 8 per cent of the world’s total over this period, a score second only to the United States which contributed 34.6 per cent. In comparison Germany contributed 7.0 per cent, Japan 7.3 per cent, France 5.2 per cent, Canada 4.5 per cent and Italy 2.7 per cent.

Citations provide a measure of the visibility and importance of papers in the eyes of the author’s peers. The United Kingdom does rather well in this respect, with a 9.1 per cent share of citations, again second only to the United States, which has a share of 49 per cent. However, the simple comparison of citation shares may not do justice to smaller countries, as, although they may contribute only a relatively small number of publications compared with the G7 countries, what they do contribute is, in some cases, highly cited.

A possible way to overcome this is to examine the relative citation impact of each subject area. This is defined as the share obtained by a country of world citations in a particular subject area, divided by that country’s share of world publications in the same area. A relative citation impact of greater than 1 indicates that citations in that discipline or area for that country are higher than the world average. It is a measure of both the impact and the visibility of a country’s research, as disseminated through publications, and gives some indication of the quality of the average paper. Table 1 shows the relative citation impact for a number of countries. As can be seen, the United Kingdom is in fifth place, behind the United States, Switzerland, Sweden and Denmark.

Table 1. Relative citation impact for selected countries

Country	Relative citation impact (RCI)
United States	1.42
Switzerland	1.38
Sweden	1.24
Denmark	1.16

United Kingdom	1.14
Netherlands	1.10
Canada	1.00
Australia	0.97
Norway	0.91
Finland	0.90
France	0.87
Germany	0.86
New Zealand	0.80
Japan	0.78
Italy	0.75
Philippines	0.58
Chile	0.55
South Africa	0.54
Thailand	0.52
Mexico	0.51
Hong Kong	0.51
Indonesia	0.47
Malaysia	0.35
Taiwan	0.35
Singapore	0.33
South Korea	0.32
People's Republic of China	0.27
India	0.27

Source: Author.

A further valuable measure of scientific strength in individual research areas can be obtained by ranking all countries in terms of shares of world publications, shares of world citations, and relative citation impact. In our analysis, we found that the United Kingdom comes second to the United States in publication shares in 11 out of the 20 fields we considered, the lowest placing being in physics (sixth).

Table 2 shows, for each field, the top five countries by citation shares and relative citation impact. Our general view is that it is desirable to be in the top eight performing countries to be considered as world-class in a particular field. As might be expected, the United States ranks first by citation shares in all 20 fields considered. The United Kingdom is second in 15 out of the 20, the lowest placing by this measure again being in physics (fifth).

Considering relative citation impact data, the United Kingdom is fourth, with the United States again leading. The United Kingdom is in the top six of the world in 15 out of the 20 fields, having particular strengths in plant and animal science, agriculture, pharmacology, neuroscience, biology and biochemistry. We are relatively weaker in chemistry, ecology, engineering, physics and computer science.

Table 2. Top five countries, ranked by share of the world's citations (first column) and by relative citation index (second column), in each of the 20 fields identified by ISI

Field	Top five countries by total citations	Top five countries by relative citation index
Agricultural sciences	United States, Japan, United Kingdom, Canada, Germany	Sweden, United Kingdom , Denmark, Canada, the Netherlands
Astrophysics	United States, United Kingdom, Germany, France, Canada	United States, Switzerland, the Netherlands, Chile, United Kingdom
Biology and biochemistry	United States, United Kingdom, Japan, Germany, Canada	United States, Switzerland, Sweden, United Kingdom , Germany
Chemistry	United States, Japan, Germany, United Kingdom, France	United States, Switzerland, Israel, the Netherlands, Sweden, (United Kingdom 9th)
Clinical medicine	United States, United Kingdom, Canada, Germany, France	United States, Canada, United Kingdom , Sweden, Denmark
Computer science	United States, United Kingdom, Canada, Germany, France	Israel, United States, Switzerland, Canada, Denmark, (United Kingdom 14th)
Ecology and the environment	United States, Canada, United Kingdom, Australia, Germany	Sweden, Norway, United States, Switzerland, Australia, (United Kingdom 9th)
Engineering	United States, United Kingdom, Japan, Germany, Canada	Denmark, Sweden, United States, Switzerland, Australia, (United Kingdom 12th)
Geosciences	United States, United Kingdom, Canada, France, Australia	United States, Australia, United Kingdom , Switzerland, France
Immunology	United States, United Kingdom, France, Japan, Germany	Switzerland, United States, Belgium, United Kingdom , Sweden
Materials science	United States, Japan, Germany, United Kingdom, France	United States, Denmark, the Netherlands, Israel, Switzerland (United Kingdom 6th)
Mathematics	United States, United Kingdom, France, Germany, Canada	Denmark, Norway, United Kingdom , United States, the Netherlands
Microbiology	United States, United Kingdom, Germany, Japan, France	United States, Switzerland, United Kingdom , the Netherlands, Israel
Molecular biology and genetics	United States, United Kingdom, Germany, France, Japan	Switzerland, United States, Germany, United Kingdom , Israel
Multidisciplinary	United States, United Kingdom, Soviet Union, France, Germany	United States, Switzerland, Denmark, Sweden, Canada, (United Kingdom 6th)
Neuroscience	United States, United Kingdom, Canada, Germany, France	Sweden, United States, Switzerland, United Kingdom , Denmark
Pharmacology	United States, United Kingdom, Japan, Germany, France	Switzerland, New Zealand, United Kingdom , United States, Sweden
Physics	United States, Germany, Japan, France, United Kingdom	Switzerland, Denmark, United States, the Netherlands, Israel, (United Kingdom 10th)
Plant and animal science	United States, United Kingdom, Canada, Germany, Australia	United Kingdom , Sweden, Denmark, United States, Australia
Psychology/psychiatry	United States, United Kingdom, Canada, Australia, Germany	United States, Sweden, Denmark, United Kingdom , Canada

Notes: Countries that contributed less than 0.2 per cent of publications over the 14-year period have been excluded. Where not explicit, UK rankings have been included in ().

Source: Author.

Table 3 shows the top 12 countries ranked according to papers per person (the numbers are re-scaled so that the United States is 100), and shows France, Germany, Italy and Japan with their rankings in parentheses. The leaders on this basis are smaller, mainly European countries. The United Kingdom is eighth in the world production of papers per person (behind Canada but ahead of the United States and the other G7 countries). For citations per person, there is almost the same 12 as for publications per person but with slight reordering; the United Kingdom is eighth again, but behind the United States in fifth place and Canada in seventh.

Table 3. Measures of relative performance

Index of papers per person (top 12 countries)		Index of citations per person (top 12 countries)	
Switzerland	167	Switzerland	179
Israel	152	Sweden	125
Sweden	147	Israel	105
Denmark	127	Denmark	103
Canada	127	United States	100
Netherlands	109	Netherlands	96
Finland	107	Canada	95
United Kingdom	104	United Kingdom	88
United States	100	Finland	85
New Zealand	99	Iceland	76
Norway	96	Norway	63
Australia	93	Australia	61
France (15)	72	France (15)	51
Germany (17)	67	Germany (16)	49
Japan (19)	49	Japan (19)	31
Italy (21)	41	Italy (20)	28

Source: Author.

Table 4 considers scientific output relative to government spending on R&D (both in total and excluding defence funds) as a measure of cost-effectiveness. Taking the level of funding in 1991, we looked at the ratio of citations per pound spent in terms of government total funding of R&D and also in terms of government civil spending on R&D. There are clearly a number of assumptions underlying this analysis and yet the key message is, we believe, reasonably robust: the United Kingdom comes out as the most cost-effective of the G7 countries in terms of papers and citations related to government civil spending on R&D. It also does very well in the comparison relating papers and citations with government total spending on R&D; only Canada does better. There is a marked gap – by a factor of three or more – between the output, measured in citations per unit of civil expenditure of the top three (United Kingdom, United States, Canada) and the other four. A similar, although slightly smaller gap, is also seen if total Government expenditure on R&D is used as the scaling factor (and similar patterns are seen if papers, rather than citations, are re-scaled against spending).

Some might argue that an English language bias in the ISI database explains why the United States, United Kingdom and Canada do so much better than France, Germany, Italy and Japan. However, the same broad patterns of performance among the G7 countries are seen in Table 3, where the leaders are not the English-speaking countries.

Table 4. Papers and citations per unit of government expenditure on R&D in G7 countries

G7 countries	Papers (yearly average over 1981-1994) per £ million of government R&D spending	
	Total	Civil
United Kingdom	9.6	17.2
United States	5.0	12.3
France	3.4	5.4
Germany	4.7	5.4
Japan	6.6	6.9
Canada	13.3	14.2
Italy	3.5	3.8

G7 Countries	Citations (yearly average over 1981-1994) per £ million of government R&D spending	
	Total	Civil
United Kingdom	93.3	168.2
United States	60.0	148.7
France	25.2	39.4
Germany	34.7	39.0
Japan	43.5	46.1
Canada	113.7	121.4
Italy	22.5	24.4

Source: Author.

At the sub-national level

Use of peer review in evaluation

All of the Research Councils conduct some form of evaluation of the research and training they support. One notable example is the Engineering and Physical Sciences Research Council (EPSRC) which is currently engaged in a comprehensive programme to establish a new framework for systematic evaluation of its research activities. The EPSRC's research portfolio is divided into discrete programme areas, and all programme areas have been invited to establish programme review teams drawn from established academic and industrial colleagues. The programme areas have been divided into relatively homogeneous themes or sub-programmes, and a common set of information and indicators on Quality, People and Exploitation (QPE) has been developed. The QPE data will be collected and appropriate overseas referees selected to help judge the quality or research excellence in world terms of the sub-programme. The outcomes of this evaluation exercise will be fed into the EPSRC's annual cycle of business planning. The exercise is in its first year.

Another clear example of the Research Councils' use of peer review in evaluation is in their management of their institutes and units. These institutes and units undertake research into specific areas that are closely associated within the core mission of the parent Council, and where there is need for concentrated effort in a centre of excellence that cannot be met through other forms of support. Such institutes/units form part of the UK science and engineering base, and are in many cases closely akin to university departments.

All Councils carry out periodic reviews of their Institutes, usually at five-yearly intervals. Institutes that fail to come up to the desired quality level in such assessments, which are largely based on peer review, are closed down, or given a limited period of time in which to effect the required improvements in quality.

The strategic need for the institutes is also examined at such occasions, to ensure that the institute is still required, and to guard against any undesirable mission drift.

An interesting facet of review mechanisms for institutes is provided by the Medical Research Council. It has a number of research “Units”, which specialise in particular areas of research. One example is the Dunn Nutrition Unit; another is the Unit of Applied Psychology. MRC units are based around a Director who has an international reputation for research in the core field of the unit. When the Director retires or otherwise leaves, the need for the continued existence of the unit is questioned rigorously. The default outcome is that the unit is closed, and the resources released are recycled to provide, for instance, for a new unit in a different area of medical science to be launched. Exceptionally, a unit may survive the loss of its Director but the unit will then be shaped around a new Director, which may well lead to a change of emphasis or overall strategic direction.

The Research Assessment Exercise (RAE)

The Higher Education Funding Councils also use peer review in evaluation: it is called the Research Assessment Exercise, which is carried out every four years or so. It involves centres within universities (often, but not exclusively departments) submitting nominated researchers from their centre for assessment, which is carried out by groups of peers, and largely based on the quality of papers authored by the nominated researchers. This system is therefore essentially based on assessment of past achievements, and is specifically orientated towards published work. The outcomes of the RAE reflect a peer-group assessment panel’s estimate of a discipline’s own research standing, against a number of subjective criteria. The quality ratings so produced are then used to inform the research funding formula for the next four years, i.e. the RAE outcomes influence the distribution of funds for research support from the Funding Councils to universities.

The RAE is a massive undertaking for universities and Funding Councils alike. It has widespread impact on universities and departments. Recently, there has been some suggestion that the RAE does not of itself lead to enhancements in the quality of research in the United Kingdom, but does encourage universities and departments to compete with each other, e.g. by bidding to attract star researchers in order to improve their record of achievement.

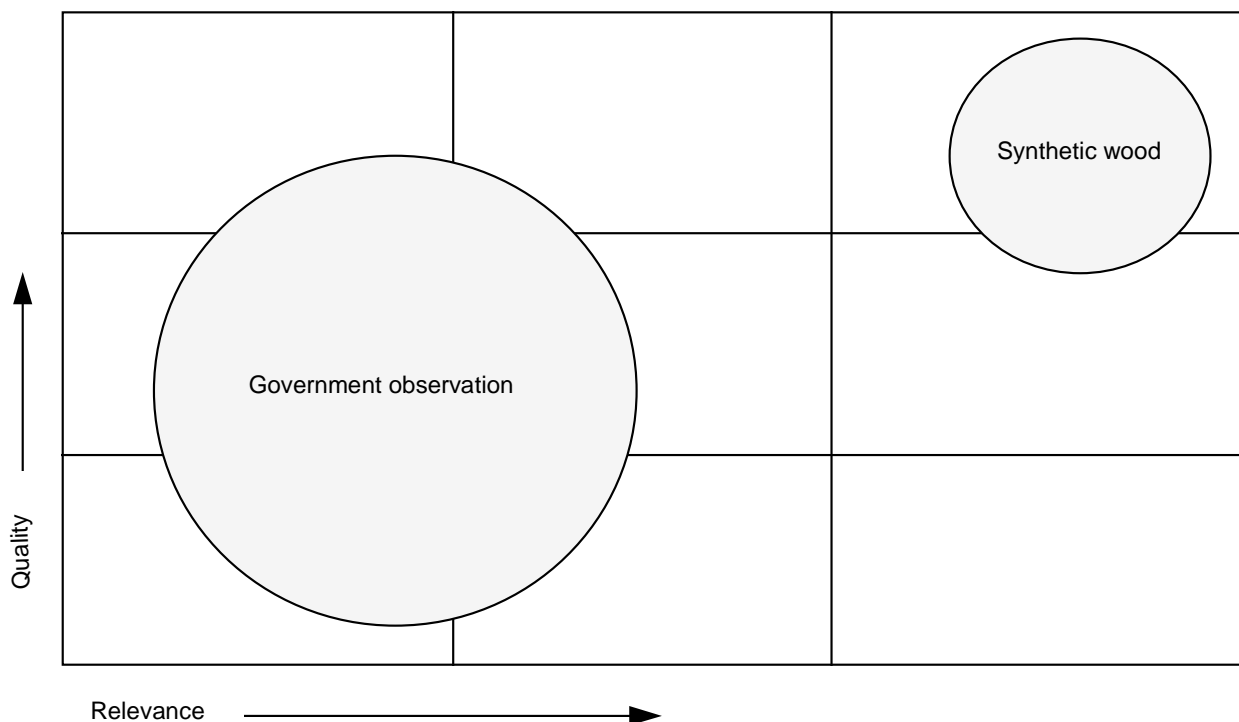
Bibliometric analysis

Several Research Councils have recently carried out reviews of the quality of science and engineering which they support, or run pilot studies to establish whether bibliometrics can readily be used to generate meaningful performance indicators. For instance, the Natural Environment Research Council (NERC) has carried out a comprehensive review of its portfolio of research, based on assessment of the quality of the research in the view of peers, and bibliometric analysis. They have also related the effort in individual topic or discipline areas to estimates of national needs, e.g. as provided by Foresight.

Representation of the results of this extensive exercise in a user-friendly fashion demands a major research programme in itself. In fact, NERC have found two-dimensional representations of the findings, using “Boston Square” diagrams to be valuable. An example of this sort of diagram is provided in Figure 1. It is desirable for research topics to occupy the top-right sector of the diagram, because this area reflects high quality and high relevance. And this is particularly important for those research activities which consume most resources, a measure of which is given by the area of the circles. It can be seen that while research into synthetic wood is apparently in good shape, research into Government Observation is less

satisfactory, being of low quality, of questionable relevance, and consuming inordinately high amounts of resources. Corrective action is urgently required.

Figure 1. Illustration of outcome of portfolio planning analysis



Note: Area of circles represents resources consumed.

Source: Author.

The Medical Research Council has just completed a pilot study of bibliometric indicators to explore the development of possible high-level indicators of research output across broad areas, to strengthen internal and external accountability and to inform policy. The pilot work covered infections research relevant to human health, and aimed to provide separate analyses for: virology; bacteriology; parasitology; prions; fungi; and vaccines and immune responses to infection. As a result of the study, the MRC satisfied itself that indicators could be developed which were seen by scientific advisors as credible and which could be shown to correlate reasonably well with peer review exercises and current opinion. The MRC now proposes to use such indicators as a policy tool.

Prior options

Finally, it may be of interest to say something about prior options. In one sense, this process is also a form of evaluation. As far as the science base in the United Kingdom is concerned it consists, in essence, of two questions:

- ◇ Is the research necessary at all?
- ◇ Need it be carried out in the public sector?

Most of the Research Council institutes have been through the Prior Options process over the last year or so. The aim has been to ensure that we obtain the best value for the money spent on research. It was determined that all the institutes were performing necessary work, and that they should remain within the public sector. However, the reviews emphasized the need for continuing work to improve efficiency in a number of cases.

Conclusions

Evaluation of basic research is an important topic, but it is very difficult to ensure that such evaluations are both timely and give messages that can be acted on with confidence. The UK approach has tended to be a combination of peer review with other indicators or analysis (e.g. bibliometrics). Perhaps the greatest value of evaluation is in helping to identify the right questions to ask, and helping to inform analysis of the answers received. And time-series data in a particular field or for a particular university or country may be more useful and relevant than attempting comparisons between the outcomes of evaluations in different areas and different fields. This is because the latter, unless done with care and insight, can lead to misleading conclusions and these in turn may lead to inappropriate action.

PART II: INSTITUTIONAL EXPERIENCES

CHAPTER 4. BIBLIOMETRIC ASSESSMENT OF RESEARCH PERFORMANCE IN FLANDERS

*M. Luwel, Ministry of the Flemish Community, Brussels, Belgium, and
E.C.M. Noyons and H.F. Moed, Centre for Science and Technology Studies (CWTS),
Leiden University, the Netherlands*

Introduction

Until the end of the 1980s, little attention had been paid to the efficiency of the Belgian research system. In contrast with most OECD countries, the evaluation of research in Belgium remained limited to peer review of research proposals. However, stagnating expenditures on higher education on the one hand, and the growing number of students on the other hand, both constituted a serious problem for science policy, and limited the possibilities for financing scientific research. Neither the institutions themselves nor the government paid much attention to investigating research carried out in various sectors. Belgium's rather complicated institutional structure was undoubtedly the determining factor behind this quite peculiar situation. The constitutional reform of 1988 was a turning point, making evaluation a primary issue on the science policy agenda of the Flemish Government.

In Belgium, universities receive a basic allowance in order to organise education and to provide basic research facilities. In spite of a significant increase in the number of university students, the basic allowance had been decreasing in real terms since 1975. Partly as the result of the rigid personnel structure, the allocation of funds for academic research stagnated. In this situation, evaluating research became imperative.

The Flemish law concerning universities, approved in 1991, obliged these institutions to develop a policy regarding quality assessment in exchange for a much greater autonomy. They had to set up procedures for a systematic review of the quality of all their activities and to report their conclusions to the government. The parliament even gave the government the authority to independently conduct a quality control review and to impose (financial) sanctions if universities were to fail to do so or if certain standards were not met.

As in most industrialised countries, at the end of the 1970s the Belgian Government became more aware of the importance of a strong R&D potential in order to sustain economic development. Although the total amount of Belgian public expenditures for R&D remained roughly constant at 0.5 - 0.6 per cent of the GDP during the 1980s, which is rather low compared with other developed countries, a series of initiatives were taken to strengthen the collaboration between industry and universities and to stimulate research in areas considered critically important in the future. For disciplines such as biotechnology and artificial intelligence, specific programmes were set up and managed by the Ministry for Science Policy. Moreover, public authorities actively stimulated universities to participate in international R&D programmes, such as the EU Framework Programme.

In this paper we discuss instruments based on bibliometric methods which were developed and applied by Flemish universities and public authorities in the evaluation of scientific research. We focus on the policy

background and policy implications of the application of these bibliometric tools. For technical-methodological issues we refer to other, recently published articles.

Bibliometric analysis

Bibliometrics involves the quantitative analysis of scientific literature. Bibliometric indicators are used as tools in two main types of analysis: the analysis of the structure and evolution of scientific sub-fields; and the assessment of scientific or technological output.

In the bibliometric analysis of the *structure and evolution of scientific sub-fields*, the first step is normally to select a set of relevant scientific documents covering an area of scientific research. The next task is the extraction of cognitively significant terms from the selected documents. The relatedness of terms is measured through their co-occurrences, i.e. the number of times two terms appear together in the same (segment of a) document. From an analysis of these co-occurrences, the terms – and consequently the publications containing these terms – are often structured by means of a two-dimensional representation. The evolutionary aspect is studied by analysing time series of data.

The *assessment of scientific output* involves the calculation of indices indicating the production, productivity, or impact of research groups. The production is measured through the number of publications published by scientists in a group. The productivity measure relates this number of publications to the research capacity of the group, which is normally expressed by the number of full time equivalents spent on scientific research. Finally, the impact is indicated by indices based on the number of times the publications are cited in some 3 500 international scientific journals covered by the Science Citation Index (SCI), produced by the Institute for Scientific Information (ISI) (Garfield, 1979).

Impact and *scientific quality* are not identical concepts (Martin and Irvine, 1983). Impact is one aspect of quality, and relates to the size of the response to a certain part of the research project as reflected in cited references in scientific articles. Bibliometric indicators are not meant to replace peer expertise, but rather to serve as a support tool in evaluation procedures (van Raan, 1993).

In the bibliometric assessment of *technological output*, data derived from patents play an important role. An interesting research topic is the study of the interface of science and technology. Citations in patents to other patents or to scientific publications, and given by the inventors or the patent examiners, are relevant sources of information. In this paper, however, we will *not* discuss the technological aspects and the potentialities of patent data, but rather focus on the scientific part in the modern R&D system.

Assessment of university research performance

The Flemish universities have to strongly justify their use of public funds and, even more so, their demand for additional support for academic research. They have the legal obligation to conduct every seven years a review of all their activities. At the same time, they are confronted not only with growing internal competition for their own research funds, but also with the demand from their researchers to provide them with more information about and assistance in the search for external research grants. The role of the research council and its administrative branch, namely the research co-ordination office, which was established in each Belgian university at the end of the 1970s, became more and more important. Its main task is to develop and implement research policy at the university level.

In order to develop instruments for its own research policy, the *University of Ghent* (RUG) was in 1990 the first Belgian university which decided to conduct a systematic evaluation of its research performance

on the basis of bibliometric techniques. A bibliometric study provided an assessment of the research activities at the faculties of medicine and science during 1980-1989. The model for this study was the measurement of research performance at *Leiden University* in the Netherlands, finished in 1983 (Moed *et al.* 1985). A basic characteristic of the methodology applied in this study involved the combination of bibliometric analyses and a validation by the scientists involved. The close interaction between the bibliometric analysts and the researchers subjected to the assessment also played an important part in this RUG study.

In 1991, the *Catholic University of Leuven* (KULeuven) and the *University of Antwerp* (UA) decided to commission CWTS to conduct a bibliometric analysis, applying the same methodology as the one used in the Ghent study. These studies were limited to the faculties of medicine and sciences for a number of reasons. First, the three universities had to finance this work themselves.

Secondly, the staff members at the faculties of medicine were familiar with bibliometric tools, especially at the KULeuven which has already been using bibliometric indicators for its internal management for more than 25 years. Finally, literature databases fairly adequately encompass the leading journals in the fields covered by these faculties.

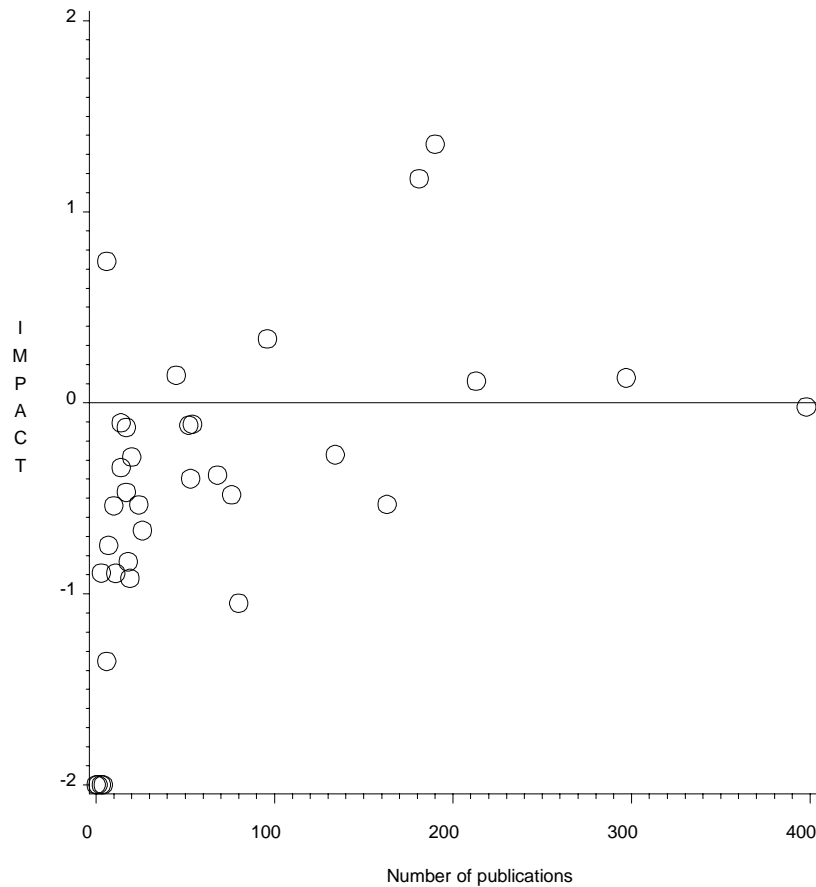
In the meantime, at the RUG, CWTS performed a study related to the other science faculties: the faculties of veterinary medicine, agricultural science, applied science and pharmaceutical science. Finally, during 1995, a more detailed analysis was made of the departments of mathematics and theoretical computer science at the KULeuven and at the UA.

For a detailed description of these five studies, we refer to Van Den Berghe *et al.* (1997). Bibliometric analyses were made of the research activities during a time period of 10-12 years of more than 400 research groups covering all these fields of science. There were close interactions with hundreds of scientists, as well as with representatives from the administrations of all ten faculties involved.

Figure 1 presents a typical outcome of the studies, related to the production and impact of all research departments in the Faculty of Sciences at the RUG. Each circle represents a research department. The horizontal axis represents the total number of articles published in SCI journals during the time period 1980-1989. The vertical axis gives the impact during the same time period of a department compared to the world citation average in the sub-fields in which it is active. Circles above the horizontal reference line represent departments for which the impact is higher than the world citation average in the sub-fields in which they are active. Figure 1 illustrates the impact of a faculty's research activities in such a way that the position of the constituent departments is still visible. Therefore, this figure represents a *synthesis* between an analysis at the level of individual departments (the *micro* level) and one at the level of a faculty (the *meso* level).

An important effect of the studies in all universities was that they stimulated discussions among scientists about the appropriateness of submitting articles to journals not processed for the SCI. In fact, several departments reconsidered and in many cases adapted their publication strategies, avoiding low impact journals as much as they could.

Figure 1. Impact and production of departments in the Faculty of Sciences at the RUG



Source: Author.

The university authorities responsible for research management concluded that the methodology applied in the studies is a valuable and very important instrument, provided that care is taken in the way it is applied, especially in the case of individual departments. The university authorities are interested in updating the studies in the future, possibly with some further refinements in the methodology. In addition, they wish to generalise the use of quantitative methods and develop instruments for the *humanities and social sciences* as well.

In fact, currently the authors of this paper are involved in a pilot project, funded by the Ministry of the Flemish Community, aimed at the development of quantitative methods for the evaluation of research performance in the fields of linguistics and law. It is expected that the role of citation analysis based on the ISI Indexes will be far less prominent in these two fields than it has been in studies related to the natural and life sciences. Important elements in this pilot study are: the development of an appropriate *classification scheme* of scholarly publications, including the assignment of *weights* reflecting the importance of each type of publication; and the collection of information from the scientists to be evaluated concerning which elements *in their own view* illustrate their national or international status.

Assessments of research fields

The Flemish Government uses several financial instruments to support R&D activities; e.g. specific programmes aimed at stimulating fundamental or basic research in sectors considered by the public authorities to be of critical importance for the economic development of the region.

In 1994, the Flemish Government ordered its Science and Innovation Administration to make, in collaboration with the CWTS, a quantitative analysis of the Flemish potential in the field of *information technology* (IT). Subsequently the results of this study were used by the experts responsible for the elaboration of the Flemish Action Plan on Information Technology, a five year programme in which industrial R&D projects were financed around a limited number of subjects. In a majority of these projects the work is done by one or more industrial corporations in collaboration with university research groups.

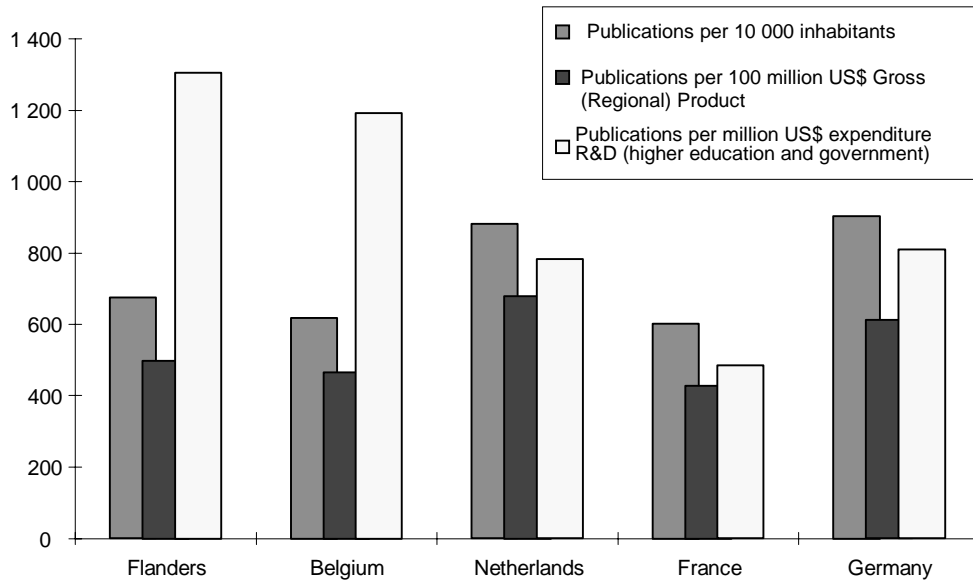
The main objective of this study was to obtain an overview of the strengths and weaknesses of Flemish R&D in the field of information technology. Data from several sources were combined in order to make the picture as complete as possible. We collected data from scientific publications, extracted from the Science Citation Index and from the INSPEC database. Moreover, data on Flemish patents were obtained from the database of the European Patent Office (EPO). Finally, we applied input statistics compiled by the OECD. The output of the Flemish IT activity was analysed in relation to the international developments in this field. We calculated the impact of the Flemish publication output in information technology and compared it to the world average. Moreover, we included data from Belgium and three European countries in the study. These three countries (The Netherlands, France, and Germany) are Belgium's neighbours and its most important trading partners. The analysis covered a period of ten years: 1983 to 1992. The results and details of the study are presented in Noyons et al. (1997).

An estimation of the Flemish activity in IT was made by normalising the output with several input indicators. From the "OECD - Main Science and Technology Indicators", the input data was obtained for the four countries included in the study. For Flanders, the data were extracted from a database with regional indicators at the Ministry of the Flemish Community. Both publication and patent data were *normalised* with the *number of inhabitants*, the *Gross National (Regional) Product*, and the relevant data on *R&D expenditures*. The results of this part of the analysis are presented in Figure 2.

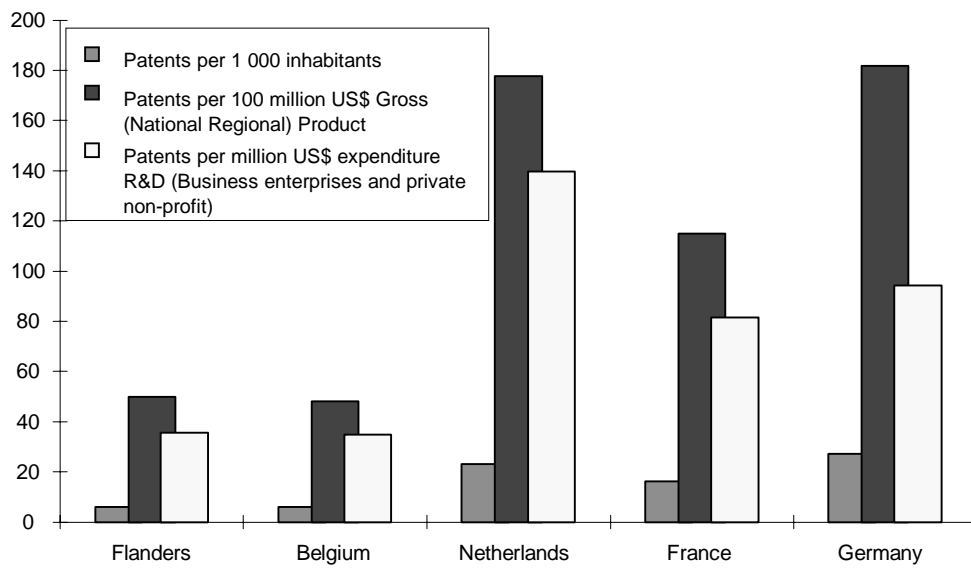
Focusing on the *publication* output, Figure 2 shows that, on the average, Flanders (and Belgium) perform at a level which is similar to that of the other European countries. On the *patent* side, however, the normalised activity is far below that of the other countries. When the publication activity is normalised with *R&D expenditures in "higher education and government"*, it is higher for Flanders and Belgium than for the three other European countries. This reflects the fact that according to the OECD statistics, public R&D expenditures, which are mainly concentrated in universities and public research institutions, are during the time period studied considerably lower in Flanders and Belgium than in the other countries.

Figure 2. Flemish publication and patenting activity weighted with OECD input statistics

Publications



Patents



Source: Author.

Assessments of public research institutions

The Interuniversity Micro-Electronics Centre (IMEC) in Louvain (Belgium) was founded in 1984 by the Flemish Government as an institute to perform scientific research which is five to ten years ahead of industrial needs. To fulfil this mission statement, IMEC has developed a strategy based on four guiding principles:

- ◇ the establishment of an internationally recognised “centre of excellence” in the field of micro-electronics;
- ◇ the performance of fundamental and strategic research in close collaboration with the Flemish universities;
- ◇ the performance of dedicated and flexible training programmes in the field of micro-electronics to both educational institutions and industrial companies;
- ◇ the reinforcement of industrial activities of companies based in Flanders.

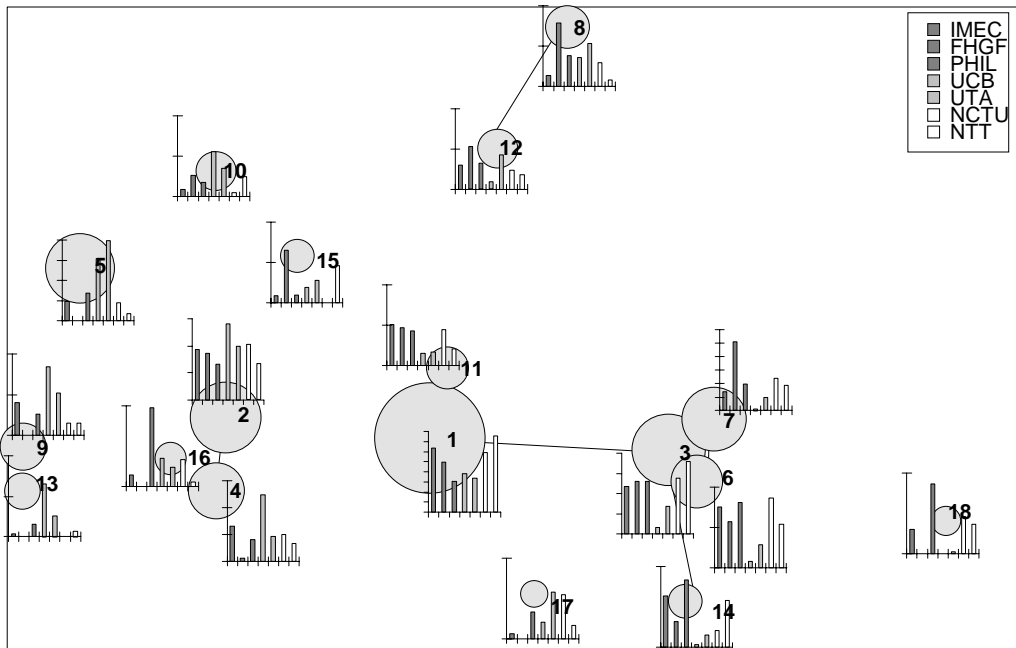
Although IMEC receives yearly from the Flemish Government a lump sum of about US\$ 30 million, it enjoys - as do the Flemish universities - a large degree of autonomy. In framework agreements between the Government and IMEC, the general strategy and objectives for a five-year period are stipulated. In the framework agreement the Government also announces a prefiguration of the block grant for each year during the considered period.

In view of the renewal of the framework agreement for 1996-2000, the Flemish Government commissioned an audit of IMEC'S activities from 1984 until 1995. This study consisted of two main parts: the first part focused on the world-wide trends in micro-electronics, and an assessment of the activity of IMEC in this field; the second part dealt with the research performance of IMEC in the field of micro-electronics as compared to the performance of selected benchmark institutes. The study provided *background material* for the Government in its negotiations with IMEC regarding the elaboration of the new framework agreement.

An interesting outcome of the study is presented in Figure 3. It relates to all publications included in INSPEC and published by IMEC and six properly selected benchmark institutes. The results of the study are still confidential. Therefore the institutes are not identified in Figure 3. This figure displays the *cognitive structure* of micro-electronics, as defined by the publications of the seven institutes covered by INSPEC. The circles represent *sub-domains*. The sub-domains are defined by sets of classification codes assigned to publications in the INSPEC database, applying the Physics Abstracts Classification System (PACS). The assignment of codes to sub-domains is established by the application of clustering techniques. The most frequently used classification codes of the field are clustered on the basis of their co-occurrences. The more two classification codes appear in the same publications, the more likely it is that they are clustered. The emerging clusters represent the sub-domains of the field. In the legend of Figure 3, to each cluster a characteristic name is given, referring to the most frequent classification codes in that cluster (sub-domain).

By this method, a cognitive structure is obtained by using the data itself, rather than a structure based on an existing classification scheme. Thus a field can be monitored from a dynamic perspective. Unexpected mergers or split-ups of traditional areas can be analysed and they shed new light on the evolution of the field as well as of the position of the actors. The research profile of an actor (e.g. a country, a university, a department) with a preference for areas with a dynamic character (unexpected merger or split-up) differs from that of an actor with a preference for more “stable” areas.

Figure 3. Structure and actors in micro-electronics



Notes: Sub-domains

- | | |
|--|--|
| 1. General micro-electronics | 10. Tele/data communication |
| 2. Circuits and design | 11. Measuring and equipment |
| 3. Materials | 12. Optical/optoelec materials and devices |
| 4. Circuit theory | 13. Control theory/applications |
| 5. Maths techniques | 14. Physical chemistry |
| 6. Liquids/solids structures | 15. Micro/electromagnetic waves |
| 7. Electronic structures/properties surfaces | 16. Radio/TV/audio; computer storage |
| 8. Optics; lasers and masers | 17. Dielectric properties/materials/ devices |
| 9. Computer theory; software engineering | 18. Supercond; magnetic properties/ structures |

Institutes:

- | | |
|------|---|
| FHGF | Fraunhofer Institut für Angewandte Festkörperphysik, Freiburg, Germany |
| IMEC | The Flemish Interuniversity Micro-Electronics Centre, Louvain, Belgium |
| NCTU | The Department of Electrical Engineering at the National Chiao Tung University, Hsinchu, Taiwan |
| NTT | NTT-LSI Labs, Kanagawa, Japan |
| PHIL | Philips Research Labs, Eindhoven, the Netherlands |
| UCB | The Department of Electrical Engineering and Computer Science, University of California - Berkeley, United States |
| UTA | The Department of Electronic and Computer Engineering, University of Texas - Austin, United States |

Source: Author.

The *relatedness* of the sub-domains, based on the number of overlapping publications, is depicted by multi-dimensional scaling. The data relate to publications published during 1992-1994. However, the structure remains stable throughout the entire period of 1988 to 1994. All the sub-domains have (more or less) the same position every year. Figure 3 reveals that the most general or basic sub-domain (*General micro-electronics*) in the centre of the map has the sub-domain 11 (*Measuring and equipment*) in its

vicinity, with an agglomeration of sub-domains in the field of materials science (3: *Materials*; 6: *Liquids/solids structures*; 7: *Electronic structures/properties surfaces*; 14: *Physical chemistry*; 17: *Dielectric properties/materials/devices*; and 18: *Supercond; magnetic properties/structures*) on the right-hand side. On the left-hand side, research topics on circuits (2: *Circuits and design*; 4: *Circuit theory*) can be found, and in their vicinity are the sub-domain 16 (*Radio/TV/audio; Computer storage*) and related topics. In the upper part of the map are sub-domains 8 (*Optics; lasers and masers*) and 12 (*Optical/optoelectronic materials and devices*).

In order to generate a general *overview of the publication activities* of IMEC and of the benchmark institutes, we labelled the relative activity in the period 1992-1994 of the investigated institutes to the sub-domains in the map. The relative activity is determined by the ratio of the number of publications of an institute in a particular sub-domain and the total number of publications by that institute. Figure 3 shows that each sub-domain has its own specific profile. On the lower right-hand side of the map (3, 6, 7, 14 and 18), there are two institutes in which activity is less prominent than in other areas. Their activity is mostly focused on the left-hand side of the map (2: *Circuits and design*, 4: *Circuit theory*, 5: *Maths techniques*, 9: *Computer theory; Software engineering*, and 13: *Control theory/applications*). One institute's activity focuses on the central area of the map.

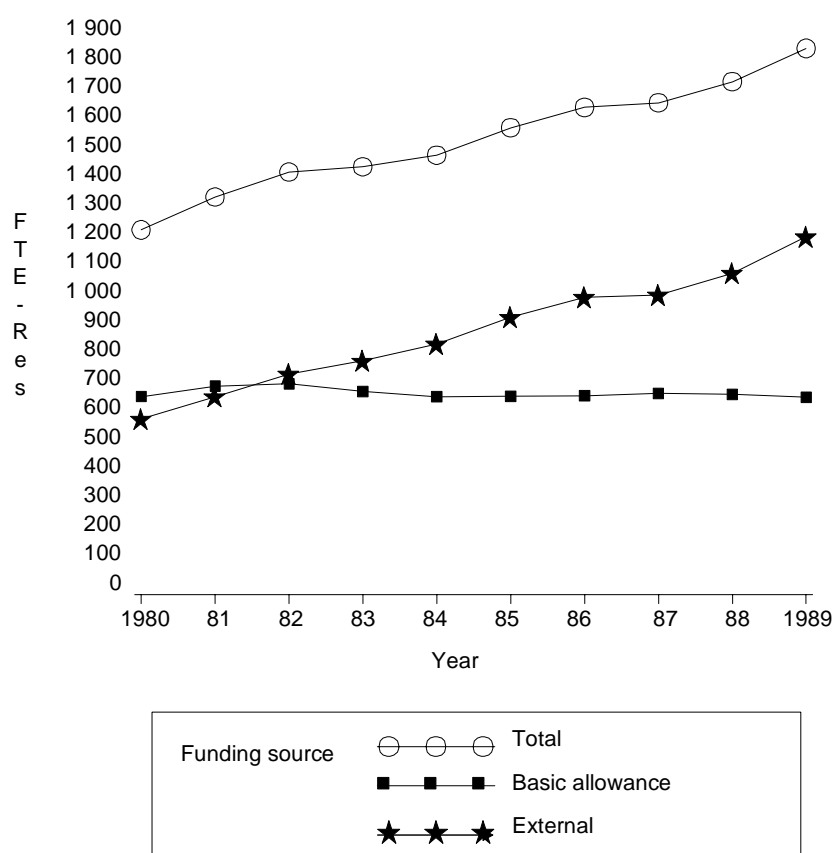
Studies on the Flemish academic science system

External funding of research projects has become more and more important in Flemish universities, while in real terms *the basic allowance*, depending upon the number of students, has stagnated or even decreased slightly over the past 15 years. Generally, during the 1980s funds for scientific research were allocated more and more on the basis of *competitiveness*, to conduct projects running for a relatively short period of time (typically, two to four years). We performed a quantitative, bibliometric analysis of research departments in the faculties mentioned above, *combining* the outcomes from previous studies, and analysing *general* patterns in the development of Flemish universities. The study focused retrospectively on the 1980s and early 1990s and addressed the following research questions:

- a) How was the development of the *scientific personnel* and the *research capacity* in the research departments involved? We focused both on the size of the research capacity as well as on its composition, particularly on the fraction of the *externally funded* research capacity and on the ratio of junior and senior scientists.
- b) How was the *distribution* of the externally funded research capacity among the research departments in the faculties involved, and were there any changes in the shape of this distribution during the 1980s?
- c) Were departments with a *high impact* in the first part of 1980-1990 able to attract *more external funding* than groups gaining a moderate or low impact?
- d) How was the development of the *productivity* - measured by the number of SCI articles per Full Time Equivalent spent on research - and *impact* - measured through citations - in the departments involved? Were there differences between departments showing a strong increase in their externally funded research capacity and groups having a moderate increase or no increase at all?
- e) How should the changes in the funding system be evaluated? What are their *positive* and *negative effects* upon the research capacity and the research performance at the universities involved?

At the level of *all departments aggregated* we found that the externally funded research capacity (RC.ext) increased during the time period 1980-1989 with about 7 per cent per year. The research capacity funded from the basic allowance decreased slightly. Figure 4 clearly illustrates this development. In this figure, it is assumed that scientists funded from the basic allowance dedicate 40 per cent of their time to scientific research. With respect to externally funded researchers, this percentage is assumed to be 100 per cent. *All* faculties showed a significant increase in RC.ext. Generally, the research capacity funded from external sources (RC.ext) was much less evenly distributed among the departments than the research capacity funded from the basic allowance. We obtained evidence that during the 1980s the *concentration* of RC.ext among departments has become even *stronger*.

Figure 4. Trend in the research capacity by funding source



Source: Author.

We classified all departments on the basis of the absolute increase of RC.ext during 1980-1989. We created four classes with departments showing a strong, a normal, or a weak absolute increase in their RC.ext, and a class with groups of departments showing no increase at all, or even a decline. The classification was such that halfway through the period 1980-1989 the total research capacities embodied in each class contained about 25 per cent of the total research capacity. However, the class showing the strongest increase in RC.ext included 18 departments, which constituted only 5 per cent of the total number of departments involved in the study. In 1989, these 18 departments allowed for 37 per cent of the

externally funded research capacity, and for 11 per cent of the research capacity funded from the basic allowance.

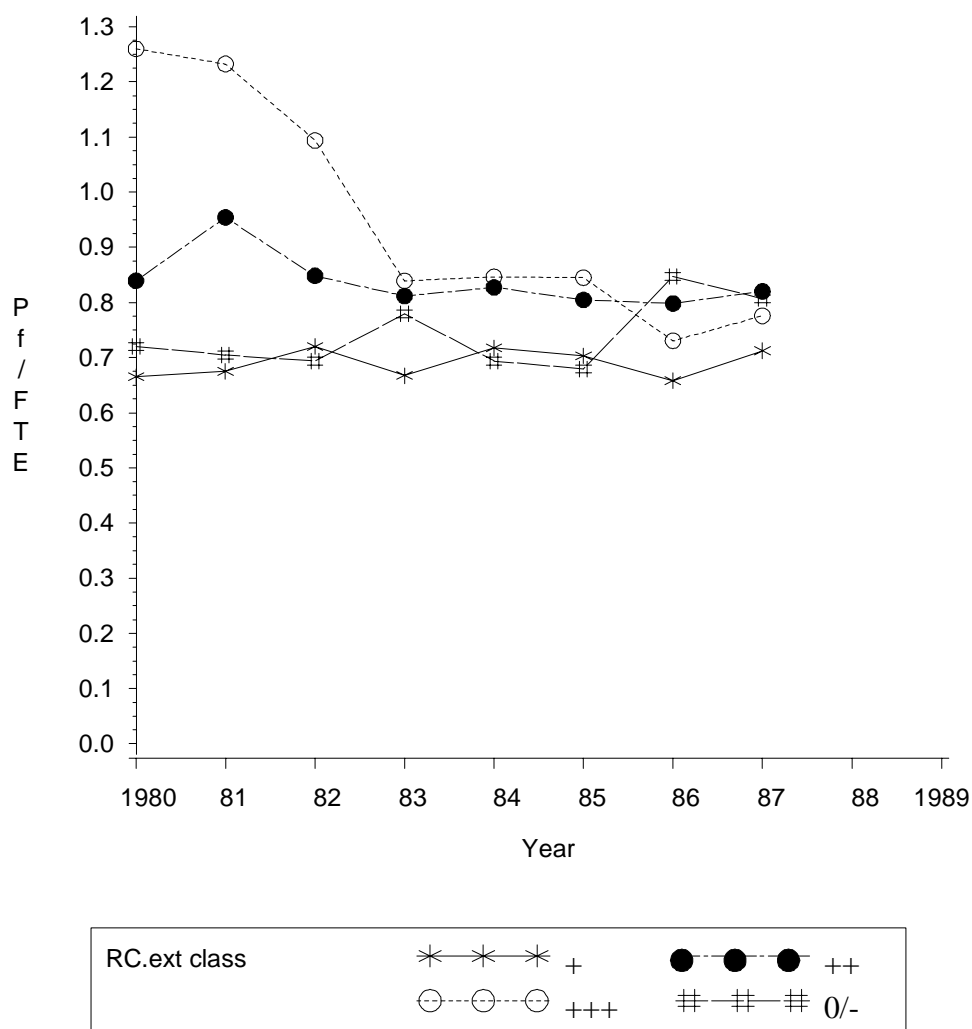
We observed a tendency in which the departments with the highest increase in the externally funded research capacity have a relatively high impact in the early years of the time period 1980-1989. This finding suggests that departments with a high international standing have profited more from external funds than groups with a lower impact. The general policy in Flanders to allocate funds for academic scientific research more and more on the basis of competitiveness seems to have been successful. International standing of research departments and their leading scientists has been a significant criterion in the allocation of external funds.

At the level of *all departments aggregated* we found that the *SCI productivity*, expressed as the number of articles in SCI journals per full time equivalent spent on research, remained more or less *constant* during the 1980s. In other words: the number of scientists in the faculties involved increased substantially, particularly due to an increase in external funding, but the SCI publication output per scientist (or more precisely, per full time equivalent spent on research) remained stable. With respect to *impact*, we observed a similar trend. The overall impact per article or per full time equivalent spent on research in the faculties involved has not increased during the 1980s.

However, we found that the SCI productivity of the class with departments showing the highest increase in the externally funded research capacity, decreased significantly during the period 1980-1989, from 1.3 SCI articles per FTE in 1980 to a value of 0.8 in 1989. This is illustrated in Figure 5. In this figure, the class of departments showing the highest increase in RC.ext is labelled with the symbols “+++”, and the class with departments with a normal or weak increase of RC.ext with the symbols “++” and “+”, respectively. The class of departments showing no increase in RC.ext or even a decrease is labelled as “0/-“. We also found that in the class of departments with the highest absolute increase in external funding, the *ratio of the number of junior and senior scientists* increased from 1.6 in 1980 to 3.9 in 1989. This ratio of 3.9 in 1989 is an “overall” ratio for all departments in the class aggregated. For some departments it was considerably higher.

The strong increase in this ratio reflects the phenomenon that external funding related to rather short term projects or programmes, and has lead to an increase of the temporary or junior scientists, while the number of permanent or senior staff members who have the duty to supervise the junior researchers remained constant or declined slightly. This phenomenon may at least partly be responsible for the observed decline in the SCI productivity of the departments showing the strongest increase in their externally funded research capacity. Evidently, the positive trend in the ratio junior/senior scientists cannot continue for ever. Our findings point towards the problem that if the trends we identified continue to develop, a situation may emerge in which the basis for externally funded research activities becomes too small.

Figure 5. Trends in the SCI publication



Source: Author.

Concluding remarks

In this article we have illustrated the use of bibliometric tools in the evaluation of scientific research conducted at Flemish universities and publicly funded research organisations, and in the assessment of the scientific-technological performance of Flanders in the field of information technology. We have sketched the policy background of a number of studies conducted during the past six years in Flanders. With respect to the effects of policy-relevant studies on policy makers, a distinction can be made between “direct” and “indirect” effects of such studies. An example of a direct effect is when a policy maker refers in his or her decisions or statements explicitly to specific results or conclusions from the policy studies. Indirect effects occur when results from policy studies are used in the policy debate to raise relevant questions, clarify concepts, question assumptions or to substantiate impressions. In terms of these distinctions, the effects of the bibliometric studies on the policy debate were mostly indirect. Nevertheless, in our view the studies have provided useful information to evaluators and policy makers in Flanders.

This is especially true for the universities. No explicit use has been made of the results obtained to allocate funds. However, in the ongoing debate at the Flemish universities and among policy makers about the creation of “centres of excellence” and a stronger concentration of research capacity on a limited number of topics, the studies formed valuable background material.

Even without a direct use by the academic authorities of their results, the bibliometric studies put even more emphasis on the research performance of departments and scientists, in an already highly competitive system.

It turns out that the academic authorities need to formulate a clear mission statement for their university and develop, in accordance, a well-balanced managerial policy taking into account education and research as well as its social, cultural and economic role. Indeed, if no overall quality assessment system is set up, the importance of quantitative, bibliometric studies can become overemphasised, leading for example in some cases to a mild neglect of education and a partial disengagement in the training of scientists from developing countries.

SOURCES

- GARFIELD, E. (1979), *Citation Indexing – Its Theory and Applications in Science, Technology and Humanities*, Wiley, New York.
- MARTIN, B.R. and J. IRVINE (1983), “Assessing Basic Research. Some Partial Indicators of Scientific Progress in Radio Astronomy”, *Research Policy* 12, pp. 61-90.
- MERTON, R.K. (1972), “The Institutional Imperatives of Science” in *The Sociology of Science*, B.S. Barnes (ed.), Penguin, Harmondsworth.
- MOED, H.F., M. LUWEL, R.E. DE BRUIN, J.A. HOUBEN, H. VAN DEN BERGHE, and E. SPRUYT (1997), “Trends in research input and output at Flemish universities during the 80’s and early 90’s: a retrospective bibliometric study”, to be published in: *Proceedings of the 6th International Conference on Scientometrics and Informetrics*, held in Jerusalem, 16-19 June 1997.
- MOED, H.F., W.J.M. BURGER, J.G. FRANKFORT, and A.F.J. VAN RAAN (1985), “The Use of Bibliometric Data as Tools for University Research Policy”, *Research Policy* 14, pp. 131-149.
- NOYONS, E.C.M., M. LUWEL, and H.F. MOED (1997), “Combining Mapping and Citation Analysis for Evaluative Bibliometric Purposes. A bibliometric study on recent developments in Micro-Electronics, and on the performance of the Interuniversity Micro-electronics Centre in Leuven from an international perspective”, Leiden: Internal CWTS Report.
- NOYONS, E.C.M., M. LUWEL, and H.F. MOED (1994), *Information Technology in Flanders* (in Dutch), Centrum voor Wetenschappen en Technologie-Studies (CWTS), Rijksuniversiteit Leiden, Administratie voor de Programmatie van het Wetenschapsbeleid (APWB), Ministerie van de Vlaamse Gemeenschap.
- PRICE de SOLLA, D.J. (1963), *Little Science, Big Science*, Columbia University Press, New York.
- VAN DEN BERGHE, H., R.E. DE BRUIN, J.A. HOUBEN, A. KINT, M. LUWEL, E. SPRUYT, and H.F. MOED (1997), “Bibliometric Indicators of University Research Performance in Flanders”, *Journal of the American Society for Information Science*, to be published.
- VAN RAAN, A.F.J. (1993), “Advanced Bibliometric Methods to Assess Research Performance and Scientific Development: Basic Principles and Recent Practical Applications”, *Research Evaluation* 3, pp. 151-166.

CHAPTER 5. RESEARCH EVALUATION AT THE CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (CNRS) IN FRANCE

Marie-Gabrielle Schweighofer, CNRS, France

The organisation of the institution responsible for basic research in France, the CNRS, includes an advisory body for research evaluation known as the National Committee for Scientific Research.

This Committee, a vast assembly made up of almost a thousand members, has two distinctive features: firstly, the evaluation of research is based on peer review; and secondly the Committee is highly representative of the French scientific community in that the latter elects two-thirds of the members of the National Committee from its own ranks.

The following presentation analyses the role, working methods and the strengths and weaknesses of this collegial, subject-based system of science evaluation, and gives some examples of changes which have recently been introduced with a view to improving the quality and efficiency of the evaluation process.

Excluding the research carried out by universities and specialised institutes, the CNRS, created in 1939, covers all areas of science. While the CNRS currently has about 350 laboratories for its own research activities, it also provides funding for over 1 000 laboratories with which it is associated. Its 12 000 full-time researchers may work in either type of laboratory (CNRS laboratory or associated laboratory), and are assisted by some 15 000 engineers, technicians and administrative staff. From an administrative standpoint, research laboratories and workers are assigned to one of France's seven regional divisions ("départements") for scientific research.

Right from the very beginning, the National Committee swiftly became the central hub of the CNRS and an institution to which the scientific community remains highly attached. The organisation and operation of the CNRS and the National Committee are closely linked. The timetable for the Committee's meetings therefore determines that of its administration, and its recommendations and advice play a major role in the decision-making process.

Role and operation

Division of science into 40 sections

In order to carry out peer review, the entire field of knowledge has been broken down into sub-domains corresponding to individual "sections" of the National Committee. This breakdown is regularly revised to take account of developments in science and the activities of the CNRS either by modifying the number of sections, the scope of sections, or their breakdown by field of activity. Thus the 11 sections originally created when the CNRS first came into being rapidly rose to around 30 by 1950 and in 1991 was increased to the current level of 40.

The main function of the sections of the National Committee is to evaluate the activities of researchers and laboratories and, unlike similar committees in many other countries, the Committee does not evaluate research programmes. There can be little doubt that researchers at the CNRS are assessed both more systematically and more efficiently than any other category of official in the French civil service.

Evaluating the activities of researchers

The first scientific evaluation is made when a researcher applies to sit the entrance competition to the CNRS. The sections sit as eligibility panels and, following a hearing, draw up a list of eligible candidates. This list is then used by admissions panels, appointed by the CNRS executive, to draw up a list of candidates for recruitment.

Because the number of applicants far exceeds the number of posts available, this first stage in the evaluation process is highly selective. Although applicants are all qualified at doctoral level, in 1996, for example, only 9 per cent of those who applied to the CNRS were eventually recruited.

Once appointed, a researcher is assessed every two years on the basis of an activity report which the researcher draws up.

The sections have a wide-ranging influence over career progression in that they put forward the names of researchers recommended for promotion and also give opinions regarding changes of laboratory, secondment to other establishments, temporary appointments of lecturer-researchers to CNRS posts, and the award of emeritus status to researchers who have reached retirement age; they also make recommendations regarding the award of bronze and silver CNRS medals to scientists.

By virtue of this Herculean labour, the sections thus provide guidance for individual researchers at regular intervals; they also provide advice on the management of researchers' careers that is much appreciated by the administration of the CNRS.

Evaluating CNRS laboratories and associated laboratories

The CNRS establishes new laboratories for an initial period of four years following evaluation of the research project by the National Committee. At the end of this four-year period, the National Committee gives an opinion as to whether the research unit should be maintained. The Committee thus plays a central role in the creation, restructuring and closure of laboratories. The turnover in terms of the number of new units set up within any given four-year period amounts to approximately 25 per cent of the 1 350 research units operated by the CNRS.

In addition to an evaluation every four years of whether a laboratory should be maintained, the activities and results of each laboratory are often reviewed after two years by the section which can thus advise the laboratory on possible future directions for its work. Laboratories whose work falls within the scope of several sections are evaluated jointly by those sections, thereby allowing account to be taken of multidisciplinary activities.

Furthermore, the sections decide which colloquia the CNRS should fund and give their opinion on editorial policy. The sections can also make appraisals on behalf of other institutions – notably regional bodies – within their given field of scientific expertise.

Reviewing trends in science

Besides making evaluations, the National Committee also fulfils another important function. Since 1959, the sections regularly review current trends in science and the prospects for future development in science at the CNRS, in France and abroad. In particular, they identify new themes emerging in their fields at the international level, the main discoveries made, current challenges and the outlook for future developments, which they use as a basis to determine the strengths and weaknesses of French research. As part of this work, every four years the Committee publishes a report on current developments in science, which serves as an invaluable tool for the formulation of science policy at the CNRS.

This exercise in critical analysis of work in its field allows each section to place its evaluation work in a national and international context. The last report by the Committee was published in 1996.

Organisations and working procedures of sections

Section members

The last elections to the National Committee, which renews its membership every four years, were held in 1995. The electorate is extremely large, amounting to around 80 000 voters drawn from researchers working in the public sector in France, researchers employed by firms that work in close collaboration with the CNRS, and the technical and administrative staff of CNRS laboratories and laboratories associated with the CNRS.

The diversity of this electoral body – from which the National Committee derives its legitimacy – is reflected in the composition of the new assembly: 476 members drawn from the CNRS; 310 members from the higher education sector; and 54 from public and private sector research establishments. This high level of representation with regard to the research community is further enhanced by a balanced distribution of members between the Ile-de-France region (43 per cent) and the other regions in France (57 per cent).

Each of the 40 individual sections has 21 members – 14 elected members and seven appointed by the Minister responsible for research on the recommendation of the Director-General of the CNRS – making a total of 840 members.

The directors of the science departments at the CNRS also take part in section meetings, together with a representative of the Ministry responsible for university research. Furthermore, if a section feels that it lacks the necessary expertise or information, it can call upon the services of outside experts to help it reach an opinion.

At its first meeting, each section elects a chairman, who plays an important role in moderating discussions. The executive of the CNRS convenes meetings with all 40 chairmen several times a year in order to advise them of, and solicit their response to, projects concerning the CNRS. These meetings allow an on-going dialogue to be maintained with the scientific community putting forward such projects.

Moreover, the 40 section chairmen regularly meet in chairmen's assemblies to consider the general issues raised by the evaluation process with regard to the future directions for work at the CNRS and in the French and international research sectors.

Organisation of the evaluation work

In all, each section examines approximately 150 CNRS researchers a year and a similar number of candidates for appointment or promotion, as well as around 30 or so laboratories and laboratory projects.

In order to complete such a large case-load of reviews and detailed examinations, the sections sit in plenary session three times a year in meetings lasting approximately three days. The sections meet in the autumn to assess researchers and laboratories simultaneously, in the spring to consider the future career of researchers, and towards the summer to consider the recruitment of new researchers.

In preparation for these meetings, five members, who constitute the section bureau, meet several weeks before the section meeting and appoint one or more rapporteurs whose task is to analyse the scientific files of the researchers or laboratories to be assessed. In the case of laboratories, the analysis by the rapporteur often includes a visit to the laboratory. The section may also organise a hearing with the director of the laboratory.

Ex post facto evaluation and consensual appraisal

The sections endeavour to remain as objective as possible in their evaluations and to avoid making poorly substantiated judgements. To this end, the evaluation is usually carried out *ex post facto* to allow the work to be evaluated over time. Solely candidates for recruitment or projected new laboratories are assessed on the basis of their potential and research programmes respectively.

Once members have given a favourable opinion or entered a reservation each section attempts to reach a consensus, which often entails lengthy debate. A recent audit found this collective exercise to be both professional and rigorous.

Criteria on which evaluations are based and transparency

Each section clearly defines its evaluation criteria as soon as it receives a mandate. The difficulty in evaluating the activities of researchers and research units lies in the many and diverse parameters that need to be taken into account. There is no common standard that can be applied to a researcher or a laboratory. The same criteria cannot be applied to both an astronomer and a biologist, for example, to a theoretical scientist and an experimental scientist, or to a young candidate and a seasoned researcher.

The evaluation criteria can therefore be tailored to meet the requirements of the individual scientific fields concerned, although in using such criteria the sections also take account of the nature of the research and the conditions under which it is carried out. A large laboratory with numerous collaborative links at the international level does not operate under the same conditions as a small, recently established team.

There are, however, many criteria that are used on a regular basis. For example, the publication of papers in reviews with reading committees, invitations to take part in international colloquia, and collaboration at international level or with industry, are objective criteria indicating that researchers and laboratories are recognised and respected at the national and international level.

In addition to exclusively scientific references, considerations such as mobility, openness to the private sector and transfer of results, teaching activities and the dissemination of scientific knowledge are new criteria which take account of the changes in the profession of researcher at the CNRS.

There must be considerable transparency in their appraisals made by sections if the opinions of the latter are to be both credible and acceptable to the community concerned. After establishing their evaluation criteria, the sections therefore make them widely known to the scientific community concerned.

In the course of the two-yearly evaluation, the opinions of the section are sent to the researchers and laboratories concerned, and also to their scientific directorate, in the form of detailed “messages” containing the observations, comments and recommendations of the section.

Strengths and weaknesses of the National Committee

The two distinctive characteristics of the National Committee, namely the classification of research activities by subject area for a minimum period of four years and the fact that the Committee is highly representative of the French scientific community, make it a powerful evaluation tool for the CNRS and one that is recognised, well-organised and stable; by the same token, however, such characteristics foster a certain degree of conservatism and inflexibility.

Breakdown of evaluation by discipline

The main asset of a section of the National Committee lies in its overall knowledge and understanding of a given area of science acquired through analysis of current developments in that field, visits to and examination of each laboratory, and evaluation of all the researchers working in its field. Consequently, unlike other organisations which ask experts or groups of experts to evaluate a given research structure or programme, each section is responsible for evaluating all existing or planned laboratories within its field of competence. The sections are therefore able, through the *comparative evaluation* of existing projects and laboratories, to guarantee the quality of all of the laboratories operated by or associated with the CNRS.

They are also able, by virtue of their overall insight into a given discipline, to determine the interest of the research projects of laboratories within the French and global context of the field, to identify areas of duplication to be avoided and thematic and geographical synergism to be encouraged, as well as the advisability of collaborative efforts that laboratories have already undertaken or could undertake.

The breakdown by discipline does have two major disadvantages, however, in that it makes it difficult to assess work at the boundaries of disciplines or on the borderlines between different disciplines. Thus when a minority area of research detaches itself from mainstream lines of research, which are generally better represented within a section, it is difficult for such an area to gain recognition for its work and to promote its researchers.

Laboratories and researchers working in several disciplines are evaluated by several sections. This is usually not a particularly satisfactory solution in that the number of new researcher posts, promotions and new laboratories from which the sections can choose is relatively limited. In the case of promotions, for example, sections tend to give preference to a researcher whose section has full responsibility for the scientific content of his work and sole responsibility for his career.

Major trends in new and well-established interdisciplinary fields involving a large scientific community are taken into account through regular reviews of the sectional breakdown of scientific research. Further to the latest review of this breakdown, for example, nine new interdisciplinary sections were created, including the “language science” section in which both the human sciences and engineering are represented.

A committee made of scientists from the same discipline will inevitably lead to attempts being made to maintain that discipline within its current boundaries and thus to a certain degree of conservatism with regard to emerging themes and new developments. The preliminary review that precedes the drafting of the report on current trends in science, by placing French research in an international context, makes it possible to gain a better understanding of trends in science and the contribution from emerging themes or themes from other disciplines (cf. the analytical framework described below).

Creating sections by discipline is wholly unsuited to the evaluation of interdisciplinary programmes. The scientific quality of the work carried out by researchers and laboratories funded under such programmes is obviously assessed by their respective sections, but in addition an *ad hoc* committee must be set up to evaluate and monitor the performance of each programme with regard to its goals. Linkage between these programme committees and the sections concerned is ensured by appointing section members to the committees.

A body that is highly representative of the scientific community

The entire scientific community is involved in the election of the National Committee. Since those eligible for election are drawn from the electorate, only researchers working in French research establishments may be elected. However, it should be noted that 10 per cent of the CNRS research establishments are foreigners and therefore eligible for election to the National Committee. Electors from industrial research centres are very seldom candidates for the National Committee because the scientific community does not know enough about their work to elect them. Another avenue to membership of the National Committee open to researchers working outside France and in industry is nomination. However, the work this entails for section members is considerable and three three-day meetings a year, coupled with the preparation of reports on some 30 files, are powerful disincentives. On the other hand, such an investment in terms of time does allow members to gain an excellent understanding of the research carried out in CNRS laboratories and to identify opportunities for collaboration.

The National Committee is therefore not an open organisation and not enough use is made of outside experts.

The representativeness of the scientific community confers a real legitimacy on the sections in that the members of sections are recognised by researchers as being one of their own and thus qualified to pass judgement on their work. Members can also fully play their role as mediators and advisers. The choices made and opinions given, even if they are negative, are usually well received.

Changes in the working methods of the National Committee

Although there has been no change in the principles on which the Committee's work is based for the past 50 years, the quality and efficiency of the work of the sections are now key issues on the agenda for discussions between section chairmen of the sections and the CNRS executive.

Outlined below are three examples of innovations that have been introduced this year: an "analytical framework" proposed for each section to guide it in its consideration of current developments in science; a fact file summarising the work of researchers applying for a promotion; and a project to computerise the scientific files of 12 000 researchers.

Framework for analysis of current developments in the discipline of a given section

An analytical framework has been drawn up to guide each section in its review of developments in science in its own particular discipline. This framework helps sections to identify the salient features and significant developments in science in their field. This framework addresses the following issues: firstly, the development of past and future understanding in the discipline; secondly, changes in its resource base in terms of instrumentation, exchanges with other fields and the organisation of the scientific community.

The dynamics of knowledge production

- ◇ What have been the main discoveries/recent advances in the field (the relevant time scale may vary from one field to another)?

Have they resulted in reference papers (if so, which ones)?

What have been the main outcomes of these discoveries (continued work along the same lines, a change in direction, a shift in themes, etc.)?

Have there been important spin-offs from these discoveries in other fields if not in other disciplines?

Have any major lines of research failed to yield results? If so, which ones? What have the consequences of this failure been?

- ◇ Which research themes are attracting the greatest interest (as shown by movements of researchers, choice of topic by doctoral students)?

Is this due solely to scientific interest or is it attributable to other factors (“fashionable” subjects, industrial interest, social interest, etc.)?

In which themes is there a marked lack of interest?

What are the reasons for this and what implications does it have for the discipline?

- ◇ What are the main scientific debates and/or controversies in the field?

Are these new debates? Have they evolved?

- ◇ Which research themes may be described as novel or emerging?

What impact will such emerging themes have on the field in the future?

Instrumentation/simulation, technological resources

- ◇ How is the role played by scientific instruments in scientific activity in the field currently evolving?

- ◇ What are, or have been, the main impacts of this on scientific activity?

- ◇ What changes have there been in the respective use of laboratory instruments and very large scientific instruments?

Trend outlook

- ◇ What impacts could the development of interdisciplinary projects have on the field?

Organisation of the scientific community

- ◇ Have any major programmes or institutes been set up abroad which have altered the organisation or balance of the field at world level?

- ◇ Have there been major changes in the nature and scale of scientific exchanges at world level in this field?
- ◇ Have such exchanges and/or collaboration received particular encouragement recently, both at international level and in France? If so, how and/or why?
- ◇ Have the supports for the dissemination of science – primarily publishing and periodicals – undergone any major changes recently?

In conclusion – challenges and aims in the field

- ◇ Is it possible to formulate scientific objectives and/or challenges of the highest importance in this field? If so, what are they?
- ◇ Have there been any noteworthy changes with regard to these objectives?

Fact file summarising the work of researchers

A file summarising the work of each researcher eligible for promotion, and drawn up by the researcher, has been introduced. This file is distributed to all members of the section evaluating the researcher.

The purpose of this file is two-fold.

The first objective is to improve the quality of the discussions held during the meeting of the section by allowing a comparison to be made of the merits of individual candidates and thus reducing the “rapporteur” effect whereby candidates with the most persuasive rapporteurs are presented in a more favourable light.

The procedure that has been applied until now has been for the rapporteur(s) alone to receive, prior to the meeting to evaluate a researcher, a copy of the candidate’s scientific file which was then used to draw up a report on the quality of the candidate’s work and the case for promotion. Since the other members of the section had no other information available during the meeting, the discussion could only be based on the arguments advanced by the rapporteur.

By means of a fact file, summarising the main items in the file, each member of the section can check that the arguments put forward by the rapporteur are indeed valid and, where appropriate, advance other arguments. A file in which fact files for all the candidates all grouped together also allows a rapid comparison to be made of candidates in the section.

The second objective is to combat the over-zealous approach to publishing, namely to give priority to the quality of publications and not their quantity.

One of the sections in the summary file asks the researcher to choose three to five publications, depending upon the level of the researcher, from his total published work. The implicit selection criteria and the publications actually selected provide two meaningful items of information for the evaluation.

Computerisation of researchers’ files

At present, to assist in the drafting of the evaluation report on a researcher, each rapporteur receives a file containing details of the researcher’s career, his activity reports and earlier assessments made of his activity by the evaluating section. This file is a paper document.

The rapporteurs for each researcher evaluated are appointed approximately a month before the meeting. Because of the time needed to study and forward files, it is impractical to appoint more than one rapporteur. For the same reason, since several sections meet at the same time, it is often not possible, in the case of research work on interdisciplinary themes, to have several sections study the file.

Computerisation of the files will enable the two rapporteurs to make cross-evaluations and improving the flow of information between sections will facilitate the task of evaluating interdisciplinary research.

CHAPTER 6. EVALUATION OF THE BLUE LIST INSTITUTES BY THE SCIENCE COUNCIL IN GERMANY

Friedrich Tegelbeekers, Geschäftsstelle des Wissenschaftsrates (Science Council Germany), Germany

Mission, background, objectives

In 1994, the Federal Government (Bund) and the 16 States (Laender) asked the Science Council (Wissenschaftsrat) to evaluate a group of 82 research or research-oriented institutions of the so-called Blue List over a period of five years (1995-1999), on the basis of the Science Council's "Recommendations for the Reorganization of the Blue List Institutes" (1993). Until the German unification in 1990, there were 48 Blue List institutes in the 11 States. Following unification, the number of Blue List institutes increased to 82 in the 16 States, i.e. the number nearly doubled. The evaluation should ensure:

1. quality of scientific work in the Blue List institutes; and
2. flexibility within the science organisation Blue List.

What is the Science Council?

The Science Council was founded in 1957 by the Federal Government and the 11 States (Laender). Based in Köln, the Science Council is an independent science-policy advisory council which advises the Federal Government and the 16 State governments of Germany. The Science Council is required to prepare reports and recommendations on the structural development of universities, Fachhochschulen (comparable to polytechnics) and research institutes. It is not a funding agency.

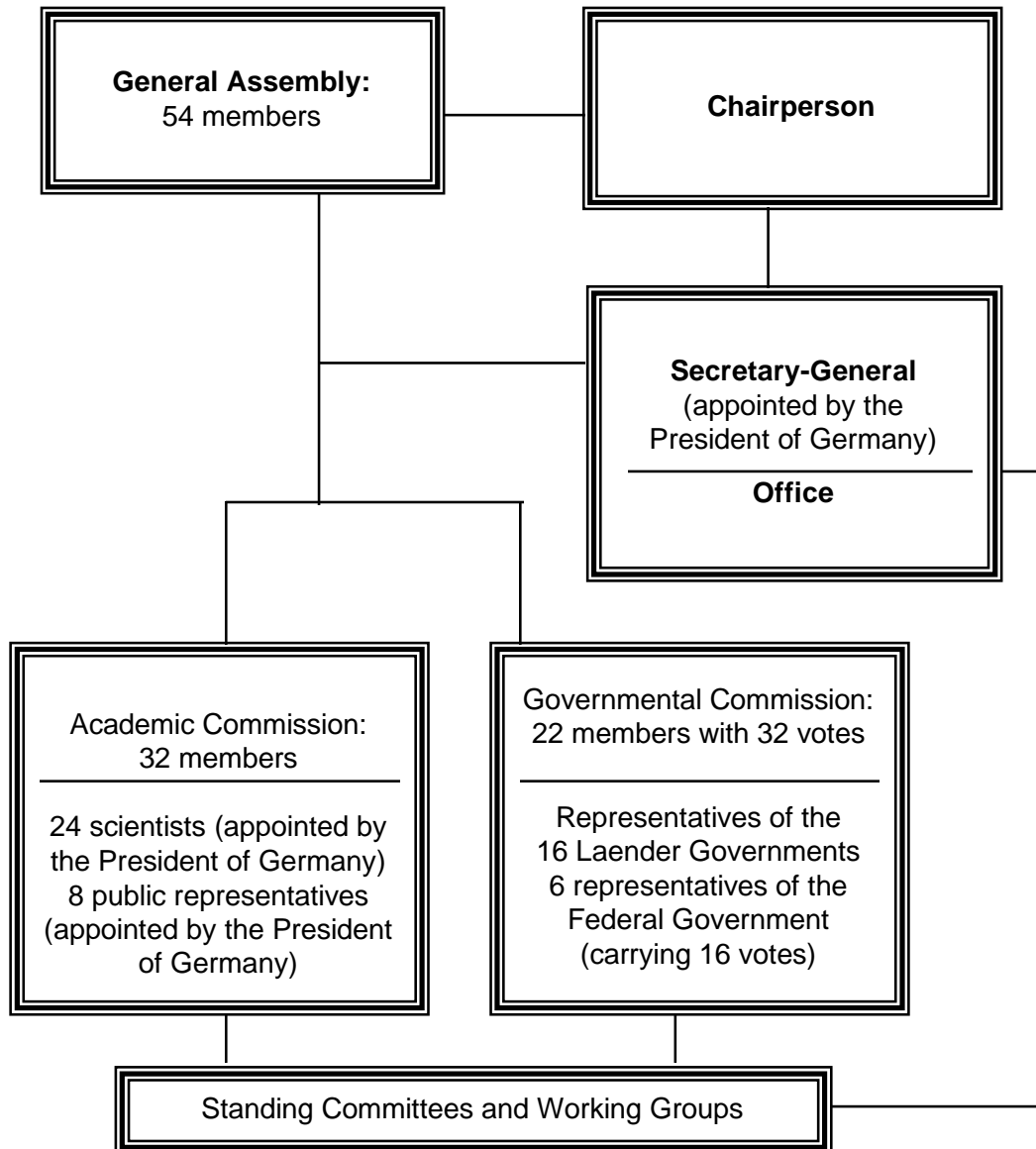
The structure of the Science Council is presented in Figure 1.

Blue List institutes: their position in the German non-university research system

The name of the science organisation refers to the fact that Blue Paper was used for the initial listing of the institutes in the 1970s.

The Blue List is a heterogeneous system of independent institutes (see Figure 2) which is made up primarily of research institutes; 16 institutes with a service function for research (for example, information centres, specialised libraries); and museums with research departments. The 82 institutes employ more than 10 000 people, among them more than 4 000 researchers. The fields covered are: humanities (16 institutes); economics/social sciences (16); life science (21); mathematics and sciences (20); and environmental science (9).

Figure 1. Structure of the Science Council



Source: Author.

Figure 2. Blue List institutes: locations in Germany



Note: Number in boxes above indicate number of institutes.

Source: Author.

The position of the Blue List institutes among the non-university research institutes in Germany in 1995 is presented in Table 1.

Table 1. Non-university research institutes in Germany (1995)

Science organisations Type of research	Institutional funding Billion DM	Positions (thousands) Research staff	Number of institutes		Federal: States share
			East	West	
Max Planck Institutes	1.5	11	10 ¹	62	50:50
Basic research		3			
Fraunhofer Institutes	0.5	6	9 ²	45	90:10 (50:50 for investment)
Short-term contract research		3			
National Research Centres	3.0	16	3	13	90:10
Strategic research		6			
Blue List institutes	1.3	10	34	48	50:50
Medium-term pre-competitive		4			

Notes:

1. Not including working groups at universities in the new Laender.
2. Not including branch offices in the new Laender.

Source: Author.

Main criteria for the evaluation of the Blue List institutes

Basic scheme

The most important criterion of the evaluation was the quality of research and service (see Figure 3). If the quality is positive, the two science policy criteria are checked. If the science policy criteria were positive, a recommendation was issued to continue joint funding. If the science policy criteria were negative, a recommendation for funding outside the Blue List (for instance, integration in a university) was issued. If quality was negative, there was no check of science policy criteria. A recommendation was issued to terminate Blue List funding.

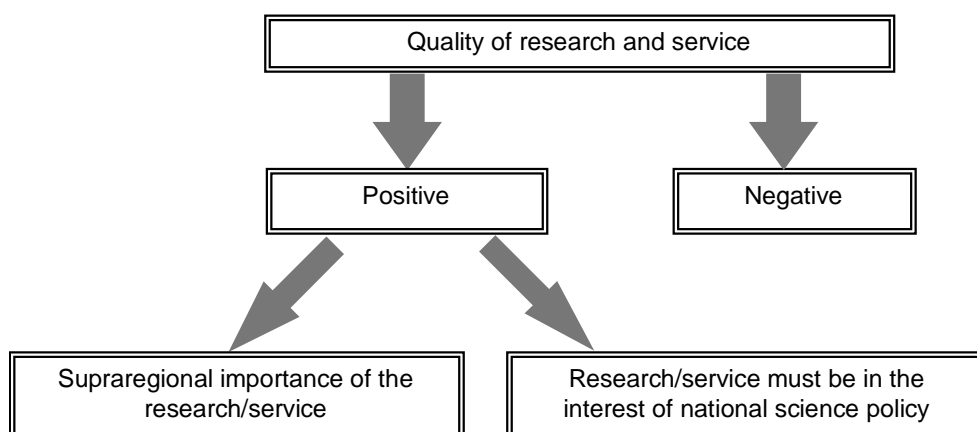
Scientific quality

The 13 criteria for scientific quality were as follows:

1. national and international integration of the institute in its main scientific field;
2. coherence of the research planning and programme;
3. qualified publications, e.g. number and quality of articles published in national and international refereed journals;

4. external funding for research projects, especially peer-reviewed funds (e.g. funds of the main German funding agency, “Deutsche Forschungsgemeinschaft”, DFG);
5. regular evaluation by a scientific advisory board;
6. qualification and flexibility of the personnel (post-graduate qualifications, up to 50 per cent fixed-term contracts);
7. co-operation with universities and research institutes;
8. joint appointments of leading academics with universities;
9. participation of academics in university teaching and promotion of young academics (doctoral candidates, postdocs);
10. number of former academics of the institute who were appointed to professorships;
11. number of academics who were invited for presentations at important national and international conferences;
12. number of academics who were invited for a research stay in institutions in other countries;
13. number of external academics who were invited for a research stay at the institute.

Figure 3. Basic scheme of main criteria for evaluation of Blue List institutes



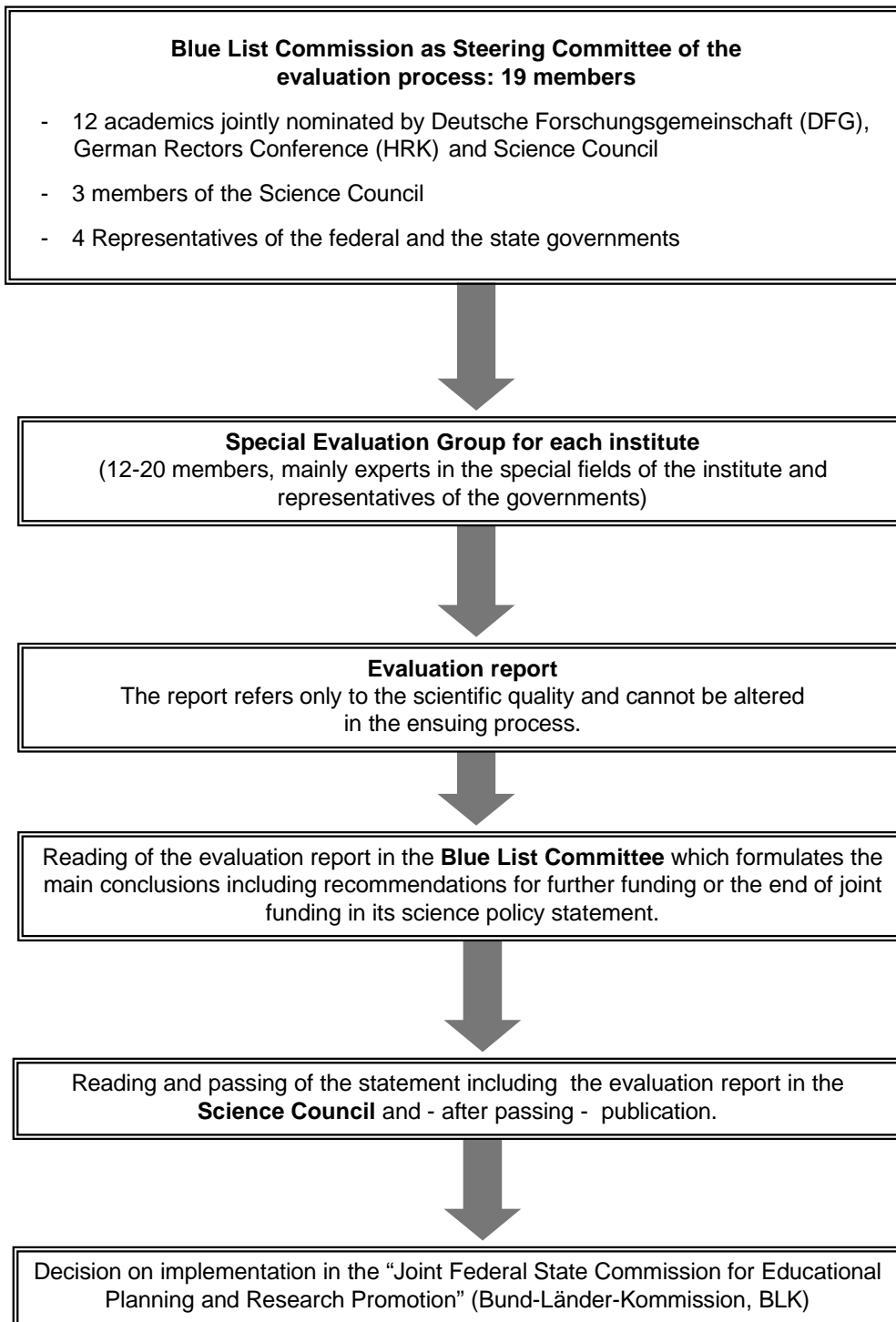
Source: Author.

Procedures and methods for the evaluation

Visiting an institute

Each institute has an on-site visit as part of its evaluation process (see Figure 4). Six months before the Science Council’s Special Evaluation Group visits the institute, the institute receives a questionnaire asking for detailed information about it.

Figure 4. Evaluation procedures and methods



Source: Author.

The on-site visit lasts between one and one and a half days. The main steps of the visit are:

1. internal discussion of the evaluation group (without members of the institute);
2. presentation of the institute by the director, leading academics and head of the scientific advisory board;
3. presentations of projects and discussions with academics in the departments in sub-groups of the evaluation group;
4. discussion of the evaluation group with the staff of the institute (without director and the leading academics);
5. discussion of the evaluation group with the president and/or representatives of co-operating universities;
6. final internal discussion of the evaluation group.

Adequacy of criteria and procedures

The criteria and procedures are the result of the Science Council's experience in evaluating the performance of research institutes since the late 1970s. These criteria and procedures are widely accepted by the evaluated institutes and the scientific community as well as by the Federal and State Governments. The procedure provides a valid basis for evaluation. It should be noted that the Science Council and its Blue List Commission do not accept the "Science Citation Index" as a valid criterion for all Blue List institutes.

First results and experiences in an on-going process

Since the beginning of the evaluation of the Blue List institutes in 1995, 17 recommendations including evaluation reports have been passed by the Science Council, of which:

- ◇ 11 institutes received a positive recommendation;
- ◇ 5 institutes received a negative recommendation;
- ◇ 1 application for funding in the Blue List was deferred.

A further 16 institutes have been visited. Recommendations and evaluation reports have not yet been passed by the Science Council.

By means of comparison, it may be important to note that during evaluation of the 48 Blue List institutes in the 1980s, the Science Council recommended terminating joint funding for two institutes only.

Since the evaluations began in 1995, findings have included:

- ◇ Institutes suffer from inflexibility (scientific programme, personnel), inadequate leadership and scientific isolation.
- ◇ Evaluation groups tend to attest quality deficits more easily than in the 1980s.
- ◇ The two-step-procedure and the inserted Blue List Commission tend to intensify the evaluation process.
- ◇ General consensus has grown in its recommendation that scarce public funds be invested where the benefits are greatest.

Problems in implementation of recommendations

The Science Council only provides recommendations. Implementation must be implemented by the Federal Government and the States (Laender) in the “Joint Federal State Commission for Educational Planning and Research Promotion” (Bund-Länder-Kommission, BLK). By means of comparison, in the 1980s the negative recommendations of the Science Council for two Blue List institutes were implemented by the BLK.

There are announcements from high-level representatives in the Federal Government and States confirming that the actual recommendations will be implemented. The process of implementation of the Science Council’s recommendations in the BLK has been initiated. In June 1997, a decision was made to stop the joint funding of two institutes; other decisions are still open. In this process, general political criteria (for instance, fiscal policy, regional policy, labour-market policy) will play a certain role too. There are clear signs that implementation will involve a very difficult process in all cases where institutes – for various reasons – have strong political support.

CHAPTER 7. RESEARCH EVALUATION AND UNIVERSITIES IN JAPAN: AN EXPERIENCE FROM THE UNIVERSITY OF TSUKUBA

Shinichi Yamamoto, Research Centre for University Studies, The University of Tsukuba, Japan

Recent changes in the research environment in Japanese universities

Massification of higher education and sophistication of advanced research

Japan has experienced enormous growth in higher education since the 1960s. The percentage of students proceeding to higher education was only 10 per cent in 1960 but had reached 45 per cent by 1995. As the percentage has grown, the enrolment has also increased from 700 000 in 1960 to 3 million in 1995.

According to this quantitative expansion of higher education, there has been a considerable diversification of the students. Universities and colleges used to be places to educate future elite and to perform research activities. But the massification of higher education has blurred this teaching-research nexus and higher education institutions must respond appropriately to the changing needs of an increasingly diverse student population.

On the other hand, research has become a more important function for some leading universities because scientific research is regarded as the engine for social and economic prosperity as well as for the advancement of curiosity-oriented inquiry. The present situation is characterised by rapid social and economic changes, including internationalisation and the shift toward an information-oriented and knowledge-based society. Scientific research is needed to expand this scope and to have a more comprehensive, interdisciplinary, and sophisticated approach.

In this situation, there is a growing expectation that universities should play a central role in contributing to scientific research and human resource development. How to advance sophisticated research under the massification of higher education has become the central policy concern which is one of the main reasons for university reform.

The growing competitive mode for resources

The Japanese higher education or university system used to be characterised by its uniformity, i.e. the similar standards of admission, the curricula, degree-granting, and the missions of each institution, as well as the organisation of the universities and their administration. For the national universities, research funding used to come from a single ministry (the Ministry of Education, Science, Sports and Culture, or “Monbusho”). This source has been used towards the basic need for faculty research as well as for the management of each institution. It has usually been distributed among faculties according to a standard formula without competition.

Recent change for improving and reforming the research system at universities, however, has made the research environment at universities more competitive. The competitive grant-in-aid has grown

tremendously and the funding from industry has also become much larger than before. In addition, the Japanese government has initiated new capital investment for research to promote specific projects on frontier and pioneering research mainly done by universities. One such investment is entitled the “Research for the future program” started by the Japan Society for the Promotion of Science (JSPS) in 1996. Funding by this programme is based on competitive proposals by university faculties.

In addition to the increase in competitive research funding, expansion of graduate education is another new trend in the Japanese universities’ research environment. Under the massification of higher education, graduate education is now expected to play various roles such as conducting advanced research, training future researchers, training professionals in various fields and so on. One of the big policy concerns is how to support graduate students and young researchers. In 1985, the Monbusho initiated a new fellowship programme called “Fellowships for Japanese Junior Scientists.” This programme supports promising young researchers, including graduate students in doctoral degree programmes, with competitive scholarship and research grants.

The university self-evaluation system and beyond

The 1990s has been a period of university reform in Japan. Established in 1987 at the Monbusho, the University Council has submitted various reports to the ministry, and the Monbusho has introduced some important policy measures under the spirit of “Remaking Universities: Continuing Reform of Higher Education”.

One of the most important measures introduced by the Monbusho is a self-monitoring and self-evaluation system set up in 1991. According to the Monbusho, continuous self-monitoring and self-evaluation are vital both as a means of revitalising universities and improving educational and research activities, and as a way of ensuring that universities fulfil their social responsibilities. In recent years there has been a growing trend toward self-monitoring and self-evaluation among Japanese universities, and many reports have been published. In fiscal 1995, reports were published by 275 universities, approximately 49 per cent of all universities.

In addition to self-monitoring and self-evaluation, some universities commission evaluations by outside experts in order to gain greater objectivity. As of October 1995, 16 universities had commissioned and published evaluations by outside experts. By seeking outside evaluations, universities can gain concrete information about the role that the community wants them to play and can identify more clearly the areas in which improvements are needed.

The need for research evaluation

Parallel with the University Council’s initiatives, the reform and improvement of scientific research systems have been discussed by science communities. The Monbusho’s Science Council has published several reports regarding the promotion of scientific research at academic institutions, in which the promotion of the quality and quantity of research was mainly discussed. The 1992 report, “The Strategies for Comprehensive Promotion of Scientific Research with a View to the 21st Century” is the base for current academic research policy by the Monbusho, i.e. the expansion of research funds, the improvement of university research facilities and equipment, the training and recruitment of young researchers, and the prioritised promotion of basic research.

The recent trend of increasing expectations for progress in scientific research and the growing cost of research activity under intensified budget constraints urge the need for introducing an appropriate

evaluation system aimed at effective use of research funding, reform and improvement of research institutions, and accountability to the public. The Council for Science and Technology began to discuss national general principles for research evaluation. In response to this, the Working Group on Research Evaluation of the Science Council (at the Monbusho) composed an interim report on fundamental perspectives on research evaluation systems based on the specific features of academic research discussed at the Science Council. The general principles elaborated in the interim report are presented in an annex to this paper.

Structural features of the University of Tsukuba regarding evaluation

Basic concepts of the University of Tsukuba

The University of Tsukuba was established in 1973 as a completely new type of national university located in Tsukuba Science City near Tokyo. The large expectation at that time for this university was that it would play the leading role in university reform by establishing free exchange and close relationships in both basic and applied sciences with educational and research organisations and academic communities in Japan and overseas, as well as by pursuing education and research to cultivate men and women with creative intelligence and rich human qualities.

To attain the highest level of education and research, the university adopted a different type of organisation than the traditional national universities such as the University of Tokyo. To ensure precise, prompt decision making, for example, the management system is specialised and centralised to reflect the opinions of each faculty. Under the leadership of the central administration headed by the president and vice presidents, the University Senate, as the highest council for university administration, the Personnel Committee, the Finance Committee, the Education Council, the Research Council and various other councils have been established to provide coherence to the university. This system is based on an idea different from that of the traditional type of university where each department has strong autonomy and the university is managed by a federation of various departments.

One of the specific features of the University of Tsukuba is the functional separation of education and research. In addition to an educational organisation called the College Cluster and School, research institutes, special project research groups and research centres have been established. The Centre for University Studies is one such research centre. Twenty-six research institutes have been established according to academic disciplines. Faculties belong to one of these institutes, conducting their individual studies in accordance with their discipline while teaching in the undergraduate and graduate courses at College Clusters and graduate schools.

Re-allocation of general university funding through the internal review process

The unique in-university research projects include individual research, joint research on common subjects, and research of promising young scholars recruited from throughout the university. These projects are given priority in the distribution of research funds to promote the advancement of academic research. The item, number of applications, and adoptions in fiscal 1996 are listed in Table 1.

Table 1. Distribution of research funds

Item	Applications	Adoptions	Funds distributed
Special research grant	3	1	up to Y 30 000 000
Research grant A	30	9	Y 10 000 000
Research grant B	384	83	Y 2 000 000
Research grant for Young researchers ¹	365	184	Y 600 000
Young researchers ²	63	63	Y 600 000

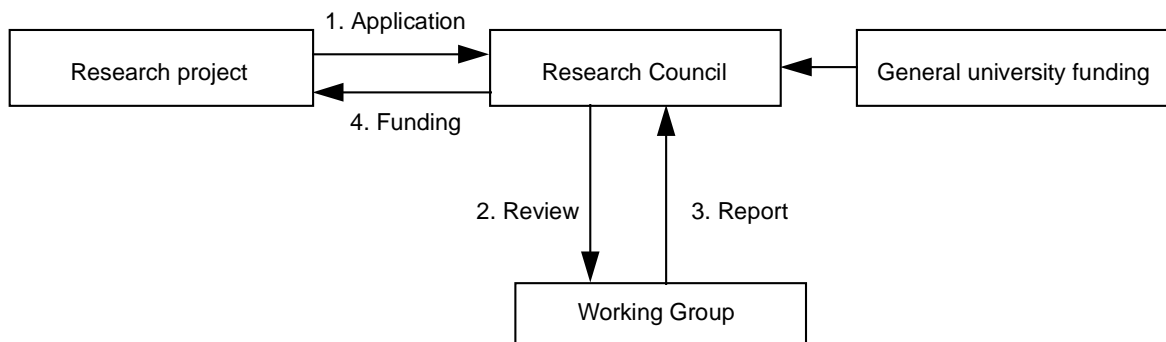
Notes:

1. Special grant to the junior faculties.
2. Special grant to the lowest rank of faculties.

Source: Author.

To provide in-university research projects with enough funding, the university keeps 8 per cent of general university funds which are otherwise distributed to all the faculties according to a certain formula base. The total amount of money available is about Y 300 million in fiscal 1996. To pick up and fund specific research projects, the university establishes the review system for the applications (see Figure 1). Small working groups composed of peer professors with a wide variety of disciplines are organised every year and each application is carefully reviewed and graded by a member of the group. The review is based on clearness of goal and method, relation of resources with the proposed research, competency demonstrated by the recent achievements of the investigator(s), adequacy of organisation of the project and so on. In some cases, the external review method is also adopted. Finally, the results are reported to the Council to be approved.

Figure 1. The system for funding in-university research projects

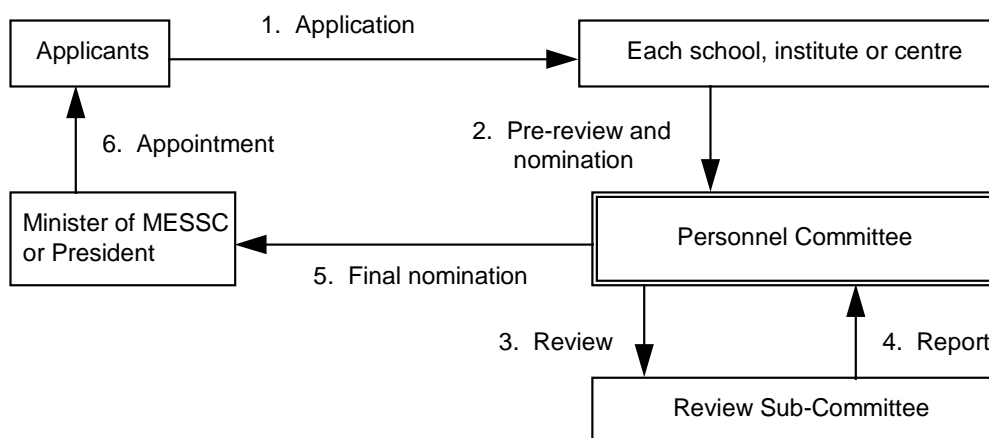


Source: Author.

The role of the Personnel Committee for the recruitment and promotion of faculties

The Personnel Committee, composed of vice presidents and ten professors elected by the Research Council and the Education Council, played the crucial role in recruiting and promotion of faculties (see Figure 2). Unlike other national universities, where each department discusses such matters and can make final decisions, the University of Tsukuba adopts central administration of personnel matters to ensure quality of faculties and due process of discussion. Thus, each recruitment, for example, must be finalised at the Committee level, although it is initially discussed and screened by a small group of peers based on an evaluation of educational and research achievements, i.e. the number of and quality of research articles.

Figure 2. The system of recruiting and promotion of faculties



Source: Author.

Recent trends for activation of research

External review programme

The University of Tsukuba has commissioned 11 external review projects since 1994. They include the external review of the Institute of Physics, College of Human Science and so on. The university plans to commission four new external review projects in fiscal 1997 including the Research Centre for University Studies.

The aims of the external review are varied and are thus not to be limited to the research evaluation. Also, the review is limited to the performance of organisations such as Institutes and Colleges and individual faculties are not the target of the review. However, the external review teams have made many suggestions and recommendations useful for reorganising the structure of research in respective fields. The following-up of each review will be the next step undertaken.

The TARA project

TARA (TSUKUBA Advanced Research Alliance), with the close co-operation of the university, the government, and industrial research, seeks to promote the most advanced interdisciplinary scientific

research, to originate basic research, and to return university research to society in the form of new industrial technologies. With this purpose, TARA was established in May 1996. Although TARA is an experimental organisation which is testing the new system for the promotion and production of new technology, its establishment implies a new direction for the improvement of research systems at the university.

TARA's most specific feature is its flexible research organisation. TARA consists of several research units called Aspect, including molecular and developmental biology, human beings in the ecosystem and so on, and each Aspect has several research projects. Selection of each Aspect and recruiting research leaders is completely based on competitive proposal and review. In addition to the review at the time of their establishment, TARA reviews its research projects every three years, and the Aspects are reviewed every seven years.

TARA incorporates the following five principles:

- ◇ Alliance with other institutions.
- ◇ Competition for the stimulation of researchers and research groups.
- ◇ Evaluation by external review.
- ◇ Priority setting for allocation of resources.
- ◇ Social contribution by return of university research results.

Regarding evaluation, TARA encourages appropriate evaluation for the operation of the system. As the university's in-house evaluation has a limited effect, TARA engages outside experts to evaluate projects and Aspects in order to promote competition and fair evaluation.

SOURCES

- THE MONBUSHO (1996), *Japanese Government Policies in Education, Science, Sports, and Culture, 1995*, Publication Bureau, Ministry of Finance.
- THE UNIVERSITY OF TSUKUBA (1996), "Outline of the University of Tsukuba, 1996-1997".
- THE UNIVERSITY OF TSUKUBA (1996), "Tsukuba Advanced Research Alliance".
- THE UNIVERSITY OF TSUKUBA (1995), *TARA News* No. 1.
- YAMAMOTO, S. (1993), "Research and Development vs. Traditionalism at Japanese Universities", *Higher Education Policies* Vol. 6, No. 2, pp. 47-50.

ANNEX

FOR A BETTER EVALUATION SYSTEM OF RESEARCH IN UNIVERSITIES

*Outline of the Interim Report of the Working Group on Research Evaluation of the Science Council,
January 1997*

Note: General principles on evaluation of research in universities are to be found in the following annex. However, detailed consideration on the evaluation of research projects for research institutes has not been included.

The Japanese Government is requested by the Science and Technology Basic Plan, which was approved by the Cabinet in July of last year, to draft a basic guideline for appropriate evaluation to be commonly used for national R&D institutions, and the Council for Science and Technology has been charged with this task. In response to this government-level effort, the Working Group on Research Evaluation of the Science Council, which is an advisory body of the Ministry of Education, Science, Sports and Culture, started discussion on how research in universities should be evaluated. This interim report describes below the current ideas of the Working Group on the basic guidelines to be used for research in universities, which need to be reflected in the discussions in the Council for Science and Technology.

Principles of evaluation for the promotion of creative scientific research

Features of scientific research in universities

Universities and inter-university research institutes (hereinafter referred to as “universities”) have as their mission to succeed to develop the intellectual assets of human kind accumulated in many fields from humanity to social science to natural science, to educate the next generations with the fruits of their scientific research, and to contribute to the development of the nation, the society, and the world.

Scientific research in universities with such a mission has its own cultural value and forms a solid foundation and driving force for the development of the nation and society, because the promotion of scientific research leads to the development of national culture and the raising of the nation’s status in the world. It also plays an important role as the intellectual and spiritual base for a variety of activities in society including the development and the creation of industry, improvement in international competitiveness, solutions to global issues, the raising of the standard of living, etc.

Scientific research in universities cannot produce good results without researchers’ free-spirited ideas and sufficient motivation to research. It is for this reason that academic freedom is guaranteed for research in universities.

Viewpoints for evaluation

The way of conducting scientific research is quite diversified. It often brings us unexpected and useful outcomes after a long period of time. Therefore we cannot promote good scientific research if we are too impatient to get visible results. So it is essential to do an evaluation of scientific research from the point of view of encouraging research activities, paying special attention to the features above.

There are two main viewpoints for evaluation of scientific research, namely, scientific significance and contribution to the society and the economy. It is appropriate, however, that we attach more importance to the former, and that we take into consideration the latter according to the field and purpose of research.

For an evaluation from the viewpoint of scientific significance, the criteria should include the level of research, the possibility of its future development, and the contribution of the research to other academic fields and disciplines. For an evaluation from the viewpoint of contribution to the society and the economy, the criteria should include the creation of new technologies, the formation of a new industrial foundation, the improvement of the standard of living, and the contribution to the development of culture. Scientific research in universities also needs to be evaluated from an educational point of view.

Evaluation of universities should be performed in light of their purpose of establishment. It should consist not only of the evaluation of the overall activities of the institution but also of the evaluation of the activities of each faculty member. One-sided evaluation should be avoided, particularly in the case of universities. Due attention must be paid to the balance and the co-ordination between research activities and educational activities.

Evaluators

Since the result of evaluation of scientific research depends heavily on the subjective judgement of the evaluators, securing a sufficient number of good evaluators is vital to impartial evaluation. At the same time, it is necessary to take measures to reduce their regular duties in their universities.

It is also important to secure evaluators with an international view and to include foreign researchers in the group of evaluators.

In the case of research that has great influence on society, it is important to listen to the views of outside experts on the social influence and significance of the research.

It is necessary to discuss the issues of evaluators' ethics, their terms and reappointment, the evaluation of evaluators themselves, and the specific system for selecting proper evaluators, etc.

Development of the supporting system for evaluation

As the institution that collects, arranges, analyses, and provides the nation-wide information useful for better evaluation, e.g. objective data that can be used as the performance indicators for evaluation, the National Centre for Science Information Systems (NACSIS) needs to be enlarged and reinforced.

A larger budget is needed at both the national and institutional levels for continuous execution of more proper and reliable evaluation.

The government and individual universities are required to secure and train specialists who help evaluators.

Accessibility of evaluation results to the public

Not only in order to secure fairness and properness of evaluation but also to obtain understanding and support from society, it is important to make the results of evaluation open to the public by various means and to receive opinions, which can be considered for future policy measures to promote scientific research and stimulate research and educational activities in universities.

From the point of view of protecting personal information, it is necessary to give careful consideration to the specific ways of sending out evaluation results to the society.

CHAPTER 8. INTERNATIONAL EVALUATIONS OF THE SWEDISH NATURAL SCIENCE RESEARCH COUNCIL (NFR)

Dorothy Guy-Ohlson, Swedish Natural Science Research Council, Stockholm, Sweden

Introduction

Pertinent background information

The Swedish Natural Science Research Council (NFR) is a governmental body which is the main funding agency for support to basic research in the natural sciences in Sweden. The support is varied: providing grants for research, initiating research projects, establishing research posts, granting scholarships, promoting international co-operation, supporting scientific publications and informing the general public about research.

The approach is essentially what is known as “bottom up” (i.e. not steered, nor restricted in any way), though some money may be reserved for priority areas selected by the NFR or chosen on the recommendation of the Ministry of Education and which naturally reflect current government policy. The subjects of the priority areas are usually of an interdisciplinary nature.

The natural science disciplines under NFR are physics, mathematics, chemistry, biology and the earth sciences. For more information about the NFR, its organisation and the sums of money involved, reference should be made to the NFR information brochure, in English. Suffice to say that the total amount of money involved means that both the government and the taxpayer, from whom the money comes, hold NFR accountable as to how it is spent. The financial accountability is carried out in the normal fashion by audit, but the scientific accountability is done through international reviews and evaluations and approximately Skr 1 million is set aside each year in the budget for this purpose.

Reviews and evaluations

The word international is taken to mean that experts outside of the Swedish academic/scientific system are invited to participate. The invited experts are usually not only renowned in their own research field, but have also played a leading role in international activities whether it be in other research funding organisations, councils of scientific societies, as an international scientific advisor, a consultant to the UN, an international journal editor, or in international correlation programmes or international expeditions etc.

Reviews of whole scientific areas (disciplines) e.g. chemistry (NFR, 1995a), biology (NFR, 1995b), mathematics (NFR, 1995c), and earth sciences (NFR, 1995d), have been carried out and published in 1995. The aim of these reviews was not only to look at the current role and status of the subjects in Sweden (even in a social context), but also to pin-point problems and make, where possible, recommendations for changes to the government, to the NFR Council and to the universities and academic system as a whole. The reviews were also to examine if there were justifiable reasons why certain aspects of research were missing in Sweden from a scientific discipline. When reviews of this kind are made they

are generally on the basis of reference groups with no individual persons, teams nor projects being named or mentioned but only the subject in its entirety being examined.

Reviews, however, do not form the main substance of this paper, but should further detail or information be of interest reference should be made to the NFR literature listed in the bibliography. Attention here, however, is focused on the evaluation of subjects, i.e. evaluation of research projects and the individual scientists or teams who have received funding to carry them out.

Limitations and aims

Concentrating on subject evaluation gives some insight into how such an evaluation is conducted in Sweden (so that a kind of general picture is obtained). Thereafter the end product and how it is used by the NFR, and by those who are evaluated, should be considered. The illustrations are drawn from the earth sciences from the recent subject evaluation of mineralogy and experimental petrology (Mao *et al.*, 1997), but the procedure there is by no means specific to the earth sciences but is applicable to the research project evaluations of other disciplines.

Practical preparatory phase

Programme committee involvement

Each discipline has a steering group, called the Programme Committee (PC). In earth sciences there are at present 18 members of the PC and each of them usually serves for a period of three years which is consecutively renewable only once, for a further three years. The members represent scientific expertise in their own sub-field of the discipline and have proven track records. Among other things it is for them to decide the rotation order of the subject/sub-field evaluations. They also define that sub-fields for subjects do develop over the years and a certain amount of flexibility is necessary to accommodate changes. They also decide which scientists should be evaluated within a particular sub-field. The PC also appoints a chairman of the Evaluation Committee (EvC) whose task it is to act as rapporteur of the findings of the EvC to them and to the NFR. This chairman is also one of their own members who is not biased in the sub-field to be evaluated, but entirely familiar with the Swedish academic and research systems and able to assist the international members of the EvC. The PC also appoints a secretary to the EvC and this secretary is usually their own PC secretary.

Questionnaire and grant holder participation

A letter is sent to the grant holders by the EvC chairman and secretary requesting them to supply names of suitable reviewers. They may suggest as many international evaluators as they wish, providing these experts fulfil certain conditions.

When the composition of the international EvC is finalised by the PC more detailed instructions and a questionnaire are sent to the grant holders (Bridgewater *et al.*, 1993). The time period to be evaluated is usually the previous five to six years and a selection of ten reprints may be submitted to illustrate the research projects to be evaluated.

Planning and time-factor considerations

For the convenience of both the groups involved an attempt is made to avoid “rush-hour” times for the writing up of the reports to be submitted for evaluation and for the site visits. Thus, for example, dead-line dates for annual NFR applications etc. are avoided. In general the pattern is that the grantees have three months (June to September) to write their reports, the evaluators have them for approximately three months (October to December), the site visits take place during one week towards the end of January and the final report is published within the next three to four months. Then the circle is complete and the whole procedure starts again for the next sub-discipline evaluation.

Selection of international experts

Though the grant holders suggest names, the final choice rests with the PC. While some definite stipulated conditions are obvious, experience has shown over the years that certain recommendations are worthy of note.

- ◇ The international experts serving on the EvC must not have collaborated nor jointly co-authored papers together.
- ◇ The experts must not only cover their own speciality in the field to be evaluated, but also have a certain breadth of expertise.
- ◇ While scientific quality and experience are the prime factors when inviting experts to participate, it has been found that experts at the zenith of their career in mid-age and still actively engaged in building up their own team/department are to be recommended. They are then eager not only to give but to receive, and alert to new ideas, approaches and possibly are even interested in developing contacts after evaluation etc.
- ◇ It is always necessary to have a large number of names on the reserve list as refusals for one reason or another are common.
- ◇ The number of members serving on the expert panel depends on the number of grant holders within each component part of the sub-field to be evaluated.

Execution of evaluation

General guidelines for international expertise

Attention is focused on the fact that, though the research projects are the main elements in the evaluation, the EvC is free, and even encouraged, to comment also on structural problems, e.g. within the academic system, the age of the doctoral students, the amounts of money awarded etc. (Mao *et al.*, 1997).

Specific requests and their motivation

Under the heading, “Aspects to be covered by an evaluation” (Mao *et al.*, 1997) the points which should be specifically addressed by the international experts are given in detail. They cover, for example, such points as the scientific quality of the results obtained by the grant holders, the scientific value of the proposed projects (including possible improvements by changing the aim and/or direction of the project under evaluation), the merits of the methods used, the capabilities of the project leader and staff, the

adequacy of existing and proposed research positions, facilities and equipment and the question of increased, unchanged or decreased financial support.

Likewise under the heading, "Report of the group" (Mao *et al.*, 1997) the expert panel receives instructions as to how they should word their final assessment of the grant holders and their projects. Each grade (excellent, very good, good, fair and poor) is specifically defined and must be used consequently.

The use of the same standardised terms is necessary for priority selection. It is therefore vital that the "grades" given and corresponding terms of recommendation are used consequently throughout an evaluation.

Site visits and selected examples

Prior to the visit of the expert panel in Sweden the grantees receive suggested guidelines for the forthcoming site visits. To each head of department or departmental representative a maximum of 20 minutes is allotted for the general presentation of their department to the expert panel. Thereafter a further maximum of 40 minutes is at the disposal of each grant holder for the presentation of their individual research projects and for questioning by the experts. The form of each presentation (e.g. whether it be a lecture summary or a demonstration) is at the grant holder's own discretion.

Four examples have been selected to illustrate different evaluation results.

1. The 1997 report to the NFR is an example that represents a very good group with the potential for excellence from the Swedish Museum of Natural History. They are a young group, doing extremely well, and it is to be expected that they will make use of the results of this evaluation at the next possible opportunity for application for research funding. Here one could easily foresee the possibilities of establishing a centre of excellence in their field. This is without doubt one of the topics for discussion at future meetings of the NFR Programme Committee for Earth Sciences (PuG) (Mao *et al.*, 1997).
2. The second example is of an established senior scientist of no mean international repute, who in the late autumn of his scientific life takes a new pathway which leads to an unexpected avenue of success (Mao *et al.*, 1997). He and his department at the University of Uppsala have received considerable backing from the NFR which has resulted in the establishment of a centre of excellence in the field. Questions to be considered revolve naturally around what will happen to this centre of excellence on his retirement and how the NFR should continue with their funding of this laboratory.
3. An example of a grantee who received the grade of excellent, but had in fact had his most recent research proposals refused by the NFR is given in Mao *et al.*, 1997. It is without doubt that PuG will be paying very close attention (and remembering the words most strongly recommended by the expert evaluators) to the next application from this professor and his department at Lund University.
4. The final example (Mao *et al.*, 1997) illustrates a case where no further money has been granted, but the expert panel takes time to appreciate the potential of the project, identify problems associated with it and offer some constructive criticism which they believe could lead to considerable improvements.

Draft copy of final report

Experience has shown that it is absolutely necessary to have a first draft copy of the final evaluation report completed before the experts leave Sweden at the end of their week of site visits. It has proved extremely beneficial and expedient to all to meet each day after the site visits have been completed and after discussion write up the conclusion of the day's work.

Concluding remarks

Publication, distribution, use of final report

The length of time between the first draft report and the final published document is variable, but after thorough correction of the first draft copy most changes can be accomplished by e-mail and by fax. A second meeting of the international experts in person is seldom necessary. Five hundred copies of the evaluation report are printed and distributed to the members of the NFR, to the members of the programme committees and to the grant holders first of all. Thereafter the distribution is to the libraries and other scientific funding agencies and to the appropriate governmental departments, but the report may also be obtained on request by interested parties.

Evaluation reports are used by the NFR for future planning, priority considerations, and by the PC for the recommendation of awarding grants. The grant holders themselves cite them in their research application proposals, in their curriculum vitae and may even make use of them for salary negotiations.

Recommendations based on practical experience

The choice and composition of the expert group of evaluators is of utmost importance. Not only must personal chemistry work, but it has also been noted that panel members work best when they are "dedicated to the cause". By experience it has been found that it works well when they are at an optimal age and position in their own careers (as far as scientific experience is concerned). It is with hesitation that an age is mentioned (late forties/early fifties), but most definitely experts should be at an age when they are still interested in developing their own departments and widening their own fields of interest, and under this should be included willingness to actively read all the papers submitted to them for evaluation and not just look through them on the plane to Sweden. A good deal of energy, enthusiasm and stamina are required to do a good job of evaluation – and the hope is that those who take this assignment will also get something out of it for themselves scientifically (as no one is doing it for the financial remuneration as the honorarium isn't worth mentioning).

Last, but by no means least, no matter how hard one tries, nothing is ever perfect and flaws do exist (e.g. the recommendation of an extra post of a Ph.D. student for an evaluated grantee, but it could turn out that the person in question is unsuitable as a supervisor). So there is a need to "calibrate" the reports against the local domestic/regional knowledge of the actual research institutes etc. There also naturally exists the need to penetrate into matters of evaluating the actual evaluations e.g. this a good or bad evaluation; did the fact that no significant results were obtained mean that the money already given was wasted or was it the opposite, that much more money is actually necessary before any results can be obtained. Even NFR has started evaluating their international evaluations.

REFERENCES

- BRIDGWATER, D., B.C. BURCHFIEL, K.C. CONDIE, I.J. HAAPALA, and P.J. PATCHETT (1993), "International evaluation of endogenic geology and petrology of the lithosphere", *Report to the Swedish Natural Science Research Council (NFR, Naturvetenskapliga forskningsrådet)*, December, Stockholm.
- MAO, H.K., A. PUTNIS, D. VAUGHAN, and D. GUY-OHLSON (1997), "International evaluation of mineralogy and experimental petrology", *Report to the Swedish Natural Science Research Council (NFR, Naturvetenskapliga forskningsrådet)*, January, Stockholm.
- NFR NATURVETENSKAPLIGA FORSKINGSRÅDET (SWEDISH NATURAL SCIENCE RESEARCH COUNCIL), information brochure in English, Stockholm.
- NFR (1995a), *International review of Swedish research in fundamental chemistry*, Stockholm, ISBN 91-546-0342-0.
- NFR (1995b), *International review of Swedish research in biology within the NFR sphere of interest*, Stockholm, ISBN 91-546-0343-9.
- NFR (1995c), *International review of Swedish research in mathematical sciences*, Stockholm, ISBN 91-546-0344-7.
- NFR (1995d), *International review of Swedish research in earth sciences*, Stockholm, ISBN 91-546-0341-2.

CHAPTER 9. UNITED STATES: THE EXPERIENCE OF THE NSF'S EDUCATION AND HUMAN RESOURCES DIRECTORATE

Larry. E. Suter, National Science Foundation, Arlington, Virginia, United States

Purpose of this paper

The funding of scientific research by government agencies is carried out by selecting research projects through the review of project proposals by a group of scientists who are familiar with the area of proposed research, or peers of the principal investigator. Once projects have been selected, through the peer evaluation process, the projects may be evaluated further by the funding agency. There are several different types of evaluation of scientific research projects that must be distinguished. They differ in the amount of access required by reviewers to the project itself and in the number of projects reviewed at once as representatives of an entire "programme" of research. This paper summarises the types of evaluation activities that are carried out in the Education and Human Resources Directorate (EHR) of the National Science Foundation, and provides some examples of the types of evaluation activities that are currently underway in that directorate. These evaluation processes are then related to the new activities required by the Government Performance Review Act (GPRA) that requires all federal agencies to report on its programmes annually to Congress.

Broad types of evaluation

Peer review panels. Individually proposed projects are evaluated for selection and awarded by scientific funding agencies. They are evaluated according to published criteria by other research scientists who form a panel of reviewers of peers. The criteria include a review of prior research projects conducted by the investigator; thus, each investigator is evaluated within a history of work performed. These panels read proposals, discuss the merits of proposed research ideas and compare the proposed ideas, and recommend those that deserve award. Once the proposals have been awarded, the panels have no further responsibility for monitoring the success of the projects. The NSF programme director is responsible for recommending the final award and for monitoring the future of each project.

Project evaluation. At the NSF, especially in the Education Directorate where large education projects may be funded that seek to modify the behaviour of education institutions, continued monitoring of projects once awarded is carried out. Some evaluation of the project is the responsibility of the investigator. Funds are provided to the investigator in the award for that person to select an independent evaluator to examine the on-going research of the project and report to the investigator on those results. NSF has attempted to provide training materials for such internal project evaluation so that the quality of research projects might be improved. Each project then submits documentation to the NSF programme officer for annual continuation of multi-year projects. That documentation is evaluated by the programme officer. Further monitoring may also be conducted by the NSF programme officer by visiting sites and personally evaluating the progress of the project, by a committee of peers who evaluate projects at one-year intervals and seek information to recommend whether the project should be continued as proposed, or by outside evaluators who are hired under contract to collect information about the projects.

Programme evaluation. Individual projects are awarded funds through programmes of awards for specific scientific areas of inquiry. In the education directorate, for example, 25 programmes exist in different areas of education development. These programmes in the Education and Human Resources Directorate include such areas as Undergraduate Curriculum Development, teacher preparation, materials for teaching in science and mathematics, and research on education technology. One of the largest areas is called Systemic Reform Efforts which seek to create new approaches to the management of entire education systems such as states, cities, and local school districts. Evaluation efforts for these large programmes require extensive data collection and analysis efforts that extend over many years. Such evaluation efforts may take the form of statistical analysis, case studies, site visits, and management review by the NSF programme directors. Thus, both quantitative and qualitative methods are used for the final judgements on the success of the programme.

Portfolio evaluation. The new Government Performance Review Act (GPRA) of the US Government that was passed by Congress in 1993 and will be fully put in place by September 1997 requires additional forms of evaluation of large programmes. First, the NSF staff developed long range goals for itself, then it examined the mixture of programmes to see whether they actually were designed to address the issues of the entire goals. Thus, the portfolio of programmes must be seen as fitting a part of a larger whole. Those programmes that seem to duplicate other efforts or to lie outside of the major goals are open to redesign.

Uses of evaluation results

The evaluation community of the United States has defined two different functions for evaluation. Evaluation may be conducted as a summative evaluation, which would evaluate whether the programme achieved the intended and appropriate outcomes. Or it may be formative, which is to evaluate a programme in terms of shorter range goals so that the results of the evaluation might be used to change its course and better meet its original goals.

While NSF might conduct many evaluations for both purposes, many of our evaluations are frequently used for formative purposes to provide programme managers with useful information to redirect a project toward more successful completion. The best scientific projects are complex organisational arrangements of resources from many university research centres and individuals. Such complex arrangements frequently can benefit from outside review of others who are not intimately involved with the day to day management of the projects.

Practice of evaluation in the directorate of education and human resources

The Education directorate funds large education reform projects in elementary and secondary school districts or institutions of higher education, it also supports research efforts to improve science and mathematics education and projects on the uses of technology. The nature of the evaluation process depends somewhat on the size of the projects or programmes being examined.

Altogether, the Education directorate awards thousands of grants each year to hundreds of institutions in almost all 50 states. In addition, about the same number of proposals for grants are received but denied because of the review process. Thus, the size of the education establishment in the United States permits widespread competition among large numbers of active researchers in all areas of scientific research and provides a pool of reviewers who are not connected with existing projects and who can provide unbiased reviews of the quality of the proposed projects. The size of the research establishment facilitates objective review and evaluation of on-going research efforts.

Large education projects

Large education projects that may receive funding at significant levels (of US\$ 1 million or more) for a number of years are evaluated first by the NSF programme staff that receive detailed annual reports from the programme managers. The goals of the programme are outlined in the original programme announcement to which the principal investigators responded with a detailed proposal. In addition the principal investigators are provided with a protocol for data collection. The investigators collect the information from the school systems involved with their project and report the results to the NSF programme officer at the end of each year so that the programme officer might use it to judge the progress of the project.

Some examples of data that might be collected in elementary and secondary school systems include the student achievement levels of students and the training experiences of the teachers. In institutions of higher education for which the purpose of the project is to increase the number of minority students achieving science degrees, the number of minority students enrolled by the institution and the number graduating might be significant indicators of the progress of the programme. Such indicators might be used to determine whether management decisions are necessary to alter the direction of the project.

Other indicators of importance are evidence of whether the projects are using maximum available resources to complete their goals. For example, does the local school district receive funds from the local business community in developing its technology curriculum, is it using funds from all federal sources to affect the improvement of mathematics learning among the students in the school system.

To assist the researchers in gathering such detailed information may require the assistance of specialists in student achievement, school management, school finance or teacher certification practices. Thus, NSF has instituted a system of programme assistance which provides technical assistance to institutions that need such help when they are preparing their reports to the NSF programme officers. It is in the best interest of all persons, the project investigators as well as the NSF programme staff, that such information be collected by the best possible method.

Since many projects are carried out by scientists who have not been trained in the methods of either qualitative or quantitative data collection, new procedures are being developed to provide resources for evaluation of on-going projects. Just as in programme evaluation, the purpose of continued monitoring and evaluation of on-going projects by the principal investigators themselves is to provide a basis for improvement of the activities. NSF has awarded a grant to the American Educational Research Association to begin a training plan to increase the number of graduate students who receive training in both science and evaluation research methods by being mentored by a senior professor with experience in evaluation. Also, materials to teach scientists how to gather the necessary information to evaluate the progress of their projects is being made available on an Internet site so that all persons can locate information about the design of proper on-going evaluation of research.

Third party evaluation

Entire programmes of many projects may be evaluated by companies who specialise in data collection and analysis and who are not connected to the programmes being evaluated. These companies are selected through a competitive process managed by an evaluation staff of the Education Directorate. The programme staff develops a work statement of the programme goals and expectations of the contractor. The contractor is asked to design an evaluation process for the entire programme which is then subject to discussion and approval by a committee that includes representatives from the programme area being

evaluated. For example, if the programme was to support the adoption of laboratories for the purpose of improving undergraduate education, representatives of the physics and chemistry teaching profession would be included as advisors to the design and analysis of the evaluation report. The Education Directorate practice is to require an unbiased report by the contractor and to request that the advisory committee prepare a separate report of their views of the programme.

Programme review

NSF has conducted reviews of entire programmes every three years by asking senior researchers in the field, selected from those who are not currently funded by that research group, to visit the programme staff and review a selection of awards and to determine if the award process is being carried out fairly and whether it addresses the correct issues. The Education Directorate has initiated a new version of this process that is more intensive and involves review of reports of project activities by the senior staff of the directorate. In this Performance Evaluation Review process (PER) a large number of project investigators (perhaps 10) from a single programme are requested to present a short summary of the progress of their award and then engage in discussion of results. The PER provides an overall summary of the direction of the entire programme to the managing staff of the research programmes. By holding a review of a large number of programmes, the managers become aware of the extent of communication, or lack of it, between projects that are conducting similar types of research and they also detect weaknesses and strengths in particular approaches that are being carried out. The reports of these programme reviews are instrumental in providing the basis to the senior staff for redirecting the resources of the programme, if necessary. Such views of entire programmes are necessary to obtain a proper fit between the goals of the agency and the actual projects being performed by each major programme.

Portfolio review

Review of the make up of the entire research efforts of large programmes, such as the Education Directorate or the Mathematics and Physical Sciences Directorate, must occur at a very general level in order to answer the questions of whether the direction of research supported by the series of programmes and projects will be likely to reach the goals of the agency. Will, for example, the make up of programmes lead to new understandings of the use of technology, of the development of a scientific workforce in the next century, in the increase of scientific capabilities among the population, and in the discovery of new natural events? Scientific discovery is a fragile process based largely upon projects that were generated from the curiosity of individual researchers. The outcome of funded research cannot be predicted on the basis of the proposals received. The management questions about organising the funding resources are best focused on whether the mix of programmes will support the university-based researcher environment. Funding of scientific research, thus, requires ultimately some strategy, nearly a theory of support for scientific discovery. The new NSF strategic plan contains elements of a strategy that will be developed further and tested during the coming years.

GPRA evaluation

Beginning in September 1997, all of the NSF programmes of research will be evaluated according to the NSF Objectives that were established in the Strategic Plan for the Government Performance Results Act. These objectives shall be used to submit new requests for federal funds to the Congressional committees that oversee NSF. Thus, the primary audience for the evaluation and reporting efforts under GPRA is the Members of Congress.

The NSF goals outline the expectations for its investments. The objectives are:

1. discovery at and across the frontiers of science and engineering;
2. an integrative, dynamic coupling among research and education at colleges and universities;
3. connections between discovery and its use in service to society;
4. a diverse, globally competitive science and engineering workforce;
5. science and mathematics capabilities for all Americans that enable them to participate fully in an increasingly technological workforce;
6. a rich collection of shared resources enabling US researchers to operate at the frontier of science and engineering;
7. effective partnerships among relevant communities, organisations, and government agencies, nationally and globally;
8. a pre-eminent US presence in the global science and engineering enterprise.

Performance assessment toward these goals requires observable outputs and outcomes at various points in the continuing cycle of investments and results to assure the agency is successful in making progress toward its goals and objectives. Thus, the Education Directorate plans to use its existing evaluation structure to develop appropriate indicators of progress for each of its goals.

The Education Directorate seeks to increase the performance of the school system in understanding of mathematics and science. Thus, it has selected four of the main NSF goals for its work:

Discovery (NSF goal 1) is a goal of the education directorate because it supports research leading toward a better understanding of the education process itself.

Connections between discovery and its use in service to society (NSF goal 2) is a goal because some of the education programmes specifically seek to support mass media efforts to reach all of society.

Achieving a diverse, globally competitive science and engineering workforce (NSF goal 4) is a primary goal of many of the Education Directorate programmes to improve the quality of the teaching workforce and to provide improved materials to teachers so that instruction can be improved.

Science and mathematics capabilities for all Americans (NSF goal 5) is also a primary goal of the Education Directorate since it seeks to increase the cognitive capabilities of all students through its programmes of improved teaching practices and materials.

Currently the Education Directorate is developing specific indicators for these areas. The goal statements provide the necessary leadership in developing an understanding of what kinds of indicators should be developed. But additional understanding of the underlying processes concerning the development of capabilities for a complex society of more than 260 million persons living across thousands of miles of land, out of range of easy communication, except through mass media or the growing use of Internet connections, is needed. Such a theory of growth of science understanding is more likely to be developed with years of monitoring of the outcome of large programmes and of developing specific indicators of their progress.

