

Quality Assessment; A Challenging Task for Academies of Sciences

Pieter J.D.Drenth*

Definition and measurement

Quality seems to be a key word in many present-day deliberations on science policy and science promotion. Whether the objectives of European institutes and networks, Framework Programmes, or collaborative research projects, or the guidelines for a European Research Council, or new initiatives of ESF, or EUROHORCS are being discussed, the criteria 'quality' and 'excellence' almost unavoidably receive the highest priority. Quality has an undoubted and nearly sacred status.

A few decades ago De Groot made a similar observation. He begins an interesting discussion on the question on whether quality of education can be measured by referring to the fact that when the word 'quality' crops up, critical reasoning usually stops and contention is silenced. We all want quality: "The word is a dubious refuge to cover up the uncertainties and differences in what we actually mean" (1983, p.57).

Quality is a concept with a variety of different meanings. One should continuously demand specification if one intends to use it in scientific or policy analysis. It will then be apparent that the term is employed in different contexts and on different levels of complexity and abstraction. This is still true in the present day debate on the quality of education and research. The pre-eminence of the concept meets very little opposition, but at the same time it is often used in an undefined and multifarious fashion. An example from every day language, the expression 'We shall attempt to keep the patient alive as long as possible, but not at the expense of the quality of life', refers to a concept of quality that is far more general and also more difficult to deal with than when one speaks of the quality of a computer or help desk service. In organizational science 'quality', in the sense of 'quality control', 'quality guarantee', and 'product quality' is rather concrete and can be measured

* Pieter Drenth is emeritus professor Psychology at the Vrije Universiteit Amsterdam, The Netherlands, and President of ALLEA.

by means of the specific characteristics of the product to be manufactured, or the service to be delivered. However, the expression 'quality of work', which has been in vogue for some 30 years now, and which refers to attempts to give work a more humane character through restructuring of tasks, delegation of responsibilities, stimulating autonomy and task enrichment, has a much higher level of abstraction.

The concept of quality (and its assessment) can, therefore, be handled at a low level of abstraction, leading to simple, often only quantitative specifications and operationalisations. Quality of research can then be attached to one (or more) single, simple measure(s): the number of publications per year, or the number of patents, invitations, rewards in a given period. Likewise, in this vein quality of education can be assessed by the number of graduates per year, the relation between inflow and outflow of students, an average score on a student questionnaire evaluating a course or training offered etc.

Quality can also be handled as a complex concept. It cannot then be defined in simple terms. The final evaluation of this type of quality will be a multiple assessment of the constituent elements that can occur in varying interrelationships and interactions. In general, this approach will provide a more useful basis for understanding the complexity of multi-faceted phenomena like quality of research or education. For such an understanding, it is then necessary to clarify which elements comprise this notion of quality. In all likelihood, one is forced in such an analysis to divide the concept even further into sub-concepts.

Why quality assessment?

The general question may be posed why the concern for quality and quality assessment at universities and research institutes should be given such a high priority. The following answer can be formulated to this question:

First, and above all, there is the fact that a researcher always remains accountable. This starts with publication of experimental or analytical results and the presentation of findings to the scientific forum. Accountability vis-à-vis the scientific community, and often even the public at large is inherent in a scientific venture and provides an essential motive for the evaluation of research.

A further answer to this question stems from the recognition of the importance of scientific research for society. It is generally recognised that preservation and further advancement of the present day welfare society cannot be guaranteed without a major contribution by national research and development efforts. And this applies to the whole spectrum of sciences and humanities. To a large extent economic and social progress depends on the creation, testing and application of new ideas, insights into and theories on the nature of and connections between natural phenomena, on social, economic and legal structures as well as on the cultural and spiritual products of the human mind. And it is self-evident that maintaining quality criteria and high standards is a *sine qua non* for this significant contribution. Thirdly, scientific research has, however, become increasingly costly. The experimental equipment needed in modern life sciences, the advanced apparatus currently being used in physics and chemistry, the large scale data collection and handling in social sciences, and even the growth of high tech information in the arts and humanities all demand expensive investments in both personnel and equipment. Naturally, Research and Development cannot claim an ever-increasing slice of the national cake. In fact, research budgets of universities and research institutes in many European countries have been under pressure for quite some time now. And given these economic constraints, national Research Councils and Boards of universities, academies, and research institutes are constantly compelled to make choices and set priorities. It is evident that the quality of research is an important parameter in making such choices.

It is also evident that although for scientists quality is the major and often sole criterion for such a choice, this cannot be the case for governments and governmental institutions. A democratically elected and controlled government cannot be denied the right, or even the duty to set priorities for the distribution of resources to different scientific fields as well as the choice of societal or economic problems to be researched on the basis of deliberate socio-political considerations. In fact, this is the basic justification for a system of strategic research.

Aspects of quality

The first section pointed out that quality can either be defined and measured at a low level of abstraction, or can be seen as a complex and

abstract concept. In our view, the latter approach is more fruitful when speaking about quality of research or education. But, of course, for assessment purposes quality has to be specified and defined in sub-concepts so as to avoid subjective, global and, therefore, unreliable evaluations.

In a later section we will present an example of quality assessment, carried out by an ALLEA task force, and will then examine the issue in more detail. At this moment I plead to be relieved from the impossible task of defining quality of scientific or scholarly research in a comprehensive and generally applicable way. Let us just reiterate that quality can best be seen as a multi-dimensional concept, in which two main elements have to be incorporated: an element of competence, originality and intellectual contribution to science, to be judged by intrinsic standards, and an element of relevance or contribution to society, as judged by external, more utilitarian criteria.

Given this multi-dimensional character of research, no available single or simple performance indicator, such as the number of publications, citation indices, number of graduates, or number of patents will suffice. The more complex the phenomenon to be evaluated, the more we have to rely on judgmental processes, in which both quantitative and qualitative data are combined. In this connection it should be stressed that the well-known and often criticised peer review procedure, fallible as it may be on certain points, continues to play a significant role in quality evaluation.

In respect of the second element, the societal relevance, the following point of view needs to be emphasised. It would be most inapt to think that this element refers only to applied research and technological development, and to narrow the concept of societal relevance to practical usefulness, or, even worse, to economic utility. In fact, a sophisticated conception of relevance applies to the whole range from basic to applied research. We would like to distinguish the following four types of 'relevance':

In the first place, *intrinsic* relevance, which goes beyond the economic value and practical applicability. Research, be it in the natural sciences, in the humanities, or in the social sciences, leads to an augmentation of the body of knowledge, an intrinsically valuable and precious quality of civilisation. Raising questions on the nature and determinants of observed phenomena is a fundamental and unique characteristic of the human species and a motor for its development.

It is clear that the continuity of this scientific discourse appears to its full advantage in dialogue with the next generation. In other words, intrinsic relevance is strongly related to the educational mission of science: the transmission, re-evaluation and further development of scientific knowledge in training and education, and the enrichment of the next generation with knowledge and insight.

Secondly, *instrumental* relevance, the immediate or indirect application of research through the transformation of its findings into practical tools and instruments. Applied science research has resulted in an abundance of such devices and techniques: measurement devices, analytical techniques, tests, drugs, diagnostic aids, but also means with which to influence people, support decision making and direct or change societal systems.

In the third place, *innovative* relevance. This type of relevance refers to the contribution which scientific research can make to the creation of new knowledge and insights, which may lead to important breakthroughs in technology and development. It should be emphasised that, while instrumental relevance is often a product of what is called applied or problem driven research, this certainly is not always the case with respect to innovative relevance. Also pure, 'curiosity driven' research may turn out - sometimes unexpectedly and unintentionally - to be highly 'practical'. Pure research can lead to surprising applications, sometimes only many years later. A few well-known examples prove this point: the development of computer topography in the 50s was based on Radon theory, 40 years old at that time; the application of polymer chemistry in plastics manufacture occurred more than 30 years after its formulation, and the time lag between the development of Marconi's telegraph and Maxwell's groundwork on the transmission of electronic waves was also more than 25 years. In fact, it is this train of thought that has led to the questioning of the usefulness of the classical distinction 'basic – applied science' by many of today's science philosophers.

The fourth form of relevance can be called *contributive* relevance. Here the aim is not instrument development or technological innovation, but rather to support or contribute to decision making and policy development on the basis of scientific findings. The visibility of the scientist's involvement can vary from almost untraceable to explicitly recognisable: The scientist can actually be one of the partners in the decision-making or policy-formation process; the research results can

be used as ammunition in a discussion or debate, either to defend or to attack a certain position, or to create positive or negative attitudes with respect to a certain stance or view; or the scientist could be asked to bring in his/her expertise to the various phases of policy formation or decision making. This expertise can, of course, originate from basic research as well as from applied, problem-driven research.

What we have tried to clarify is that 'relevance for society' has many facets and can mean quite different things, and certainly refers to more than technological development. There should still be scope for research that is generated by intellectual curiosity and which aims at the augmentation of knowledge and insight as such, no matter whether this refers to phenomena in the universe, the earth system, human or animal behaviour, or cultural products like languages, and economic or legal systems. Reflection upon the nature and meaning of things, as is realised in philosophy and theology, also falls within this ken.

Criteria for quality assessment procedures

An important question to be considered in designing an assessment procedure is which criteria should be met in choosing or developing instruments or indicators of quality. The following criteria can be listed (see e.g. Drenth, 1986):

Relevance. This concerns the degree to which an index or rating instrument adequately represents the goals or the performance domain that the assessment hopes to cover. Two major questions arise here. Firstly, whether the essential elements of the intended goal are sufficiently accounted for. If this is not the case, we speak of deficiency. Secondly, whether it can be guaranteed that no aspects are included that do not belong to the area of the phenomenon to be assessed. In this case we speak of excessiveness. A rating system for quality of research in which only quantity of output is measured is deficient, and a system that includes students' ratings of teaching quality may be excessive, since quality of teaching may not be part of the defined domain 'quality of scientific research'. The relevance of a system of assessment is therefore mainly determined by the question to what extent deficiency and excessiveness have been avoided.

Validity. This refers to the empirically testable question whether an instrument or scale, intended to ascertain a certain characteristic or quality, actually does measure this characteristic or quality. Peer rating of project proposals may serve as an example. The extent to which the real quality of the proposal is rated indicates the validity of peer rating. The extent to which such ratings also reflect the reputation of the institute in which the author works, the quality of the proposal's English, or age, gender or nationality of the author invalidates peer rating.

Reliability. This notion refers to the degree to which all kinds of coincidental influences or error factors are eliminated from a given method of assessment or measurement. Ideally, two independent ratings should result in identical scores. It is clear that mostly objective, quantifiable data are more reliable than subjective, 'softer' data. On the other hand, these objective quantitative data are often more vulnerable in view of the demands set by relevance and validity. In many cases, one is forced to revert to the more subjective methods in order to guarantee sufficient relevance and validity.

Transparency. This criterion refers to the degree to which the process of making the assessment and evaluation is clear and unambiguous. Which elements constitute the final assessment, and how much weight is attributed to these elements should be transparent for those involved. The argument for this criterion is threefold. First, transparency is a general requirement that can be set for all inferences and judgements. Inferential processes and elements in evaluation should be made as explicit as possible in order to create a rational, analysable and improvable procedure. Second, assessments often have consequences: measures will be taken, sometimes with serious consequences for the organisation, personnel or finances. People affected are entitled to have insight into the process of evaluation and the weighing of various elements against one another. Moreover, in the case of unpleasant consequences for individuals, the possibility of appealing against such measures should be offered, which once again means that the grounds on which the evaluation rests should be made explicit. Third, the objective of an assessment is often improvement. And there is the learning principle that feedback for learning and improvement should be specific. Qualities and shortcomings should be presented in a clear and detailed

manner, if a modification of behaviour in a desired direction is aspired to.

Acceptability. This criterion is not totally different from the previous one. Totally non-transparent procedures are generally unacceptable for the people involved. Of course, the reverse is not necessarily true: high transparency is no guarantee of acceptability. What has been said about the importance of transparency for feedback and improvement is also true of acceptability. Change and amelioration can be best achieved if the basis on which the assessment rests is acceptable for the institution or the individual concerned. Also the problem of consequences is an important factor in this discussion. The more serious these consequences are for those involved, the greater the demand for acceptability. People are less disturbed by less acceptable criteria being applied for a scientific prize or small grant, than by such criteria being used for the evaluation of their institute or department.

Role of Academies

In my view, Academies of Sciences and Humanities (and international Associations of Academies, such as ALLEA) are appropriate organisations to shoulder the responsibility for the assessment of scientific research quality, as described above, under certain circumstances. In addition to having a platform and meeting function, the administrative responsibility for research carried out by Academy projects or in Academy institutes, and an advisory function with respect to the promotion of science, a fourth allotted task for the Academy may be an evaluative function; evaluation of individuals (prizes, scholarships, and fellowships), programmes (research programmes, and proposals for graduate schools), and institutes (research institutes within universities or other governmental organisations).

This evaluative function of an Academy can be defended on three grounds. In the first place the availability of the profuse scientific knowledge and experience within its walls and within its advisory councils and committees. Secondly, the impartiality to be expected of the Academy members. With a serious and responsible Academy no political, economic, regional or professional interest group can hope to be especially favoured in Academic judgements. Thirdly, the exclu-

sively scientific and scholarly orientation of the Academy members: it is the promotion of good science and scholarship that determines their choices and judgements.

At this point it may be appropriate to take a closer look at the distinction between this 'Academic' evaluation function and the equivalent function of National Science Foundations or Science Research Councils. A number of years ago the Dutch Academy of Sciences (KNAW) and the Dutch National Science Foundation (NWO) concluded that the tasks of both agencies could be best divided along the lines separating evaluation *ex ante* from evaluation *ex post*. NWO evaluates proposals for projects, programmes and individual activities to be financially supported in the future, and therefore works in a *prospective* context. The Academy evaluates whether and to what extent objectives set out in the past have been achieved and assesses the performance and achievements *retrospectively*. Of course, the distinction is not a 100% discriminant function, and does not preclude some of the Academy evaluations having predictive connotations, and the ratings of NWO being based on past performance as well.

A similar situation occurs at an international level. Take Europe as an example. More and more institutions, projects, programmes, and collaborative networks in Europe dissociate themselves from the national perspective and have a real supra-national, sometimes pan-European, character. The same is true of European research efforts as initiated by the European Commission (framework programmes) and, for instance, the European Science Foundation (ESF). Here the need for an 'einmalige' or periodic, independent evaluation will be felt as well. It is my view that ALLEA as the Federation of National Academies in Europe can fill the hiatus and offer its expertise for such evaluations. The ALLEA evaluation committees or evaluation committees composed by ALLEA and selected from the rank and file of Academicians, can fulfil a similar function at the European level as National Academies fulfil within their own country. It is in this context that I as President of ALLEA have asked the member Academies to nominate up to five Academicians who can be approached for such a review activity. ALLEA will then have a pool of potential reviewers to draw on should the need arise.

There is still another advantage in the creation of such a pool of reviewers. We see a growing tradition of inviting *foreign* experts on national review committees. This is to be encouraged, since this will con-

tribute to further internationalisation and counters provincialism through international benchmarking. Here, again, ALLEA members can be of service to one another through their availability as foreign expert reviewers.

An example

Late in 2001 the European Science Foundation (ESF) approached ALLEA with the request to review the structure, operations and achievements of the Standing Committees on Social Sciences and Humanities. The year before, the Standing Committees of Life and Environmental Sciences, and of Physical and Engineering Sciences, as well as the European Medical Research Councils had been reviewed by the Royal Society of London.

Terms of reference included the task to evaluate the operations of the two Standing Committees within the overall ESF structure, to consider the effectiveness, impact and recognition within the wider European scientific community, the operation of links with ESF Member organisations, to consult with relevant key respondents, and to report to the ESF Governing Council, through the Executive Board. ALLEA saw this request as an interesting challenge, and as an opportunity to acquire some experience in this area. Its proposal with respect to the design, personnel and financial requirements and the time schedule was accepted by ESF and the activities started at the beginning of 2002. The following aspects of this evaluation exercise deserve attention:

(1) *Personnel*. Given the wide theoretical and methodological differences between social sciences and humanities, it was decided to compose two review panels. Suitable chairmen for these two panels were recruited from the Dutch Academy. Both candidates accepted this responsibility. The actual review committee would exist of the combination of the two panels, while ALLEA's President would serve as a linking pin between them.

The Member Academies were then asked to nominate possible candidates for the review panels. From the suggested name pools the actual members of the two panels were selected, where possible making allowance for a proper distribution of region, discipline, and gender. The proposed composition of the panels was accepted by the ESF.

A staff member of the Dutch Academy was proposed to act as secretary of the SCH panel, and as secretary of the SCSS panel an external professional, who would also contribute expertise in survey research and science policy analysis, was employed.¹

(2) *Design*. The following design and time schedule was developed for the review:

- Until end May, 2002: Gathering of initial information through document analysis and interviews with some key informants, such as present and previous secretaries of the SCSS and SCH. Completion of the preliminary overviews of both the standing committees. A preliminary definition of the main issues.

- Until end June: Consultation with Royal Society of London and other experts on previous reviews of science panels. Development and evaluation (with both panels) of design and methodology framework. Defining populations from which samples had to be approached with questionnaires, viz. Standing Committee Members, Member Organisation Representatives, and Recipients (of ESF grant or support). For each population a specific questionnaire was to be designed. Selection of key informants for in-depth interviews.

- Until end July: Construction of questionnaires and pilot testing.

- Until end October: Distribution of questionnaires, monitoring returns. Preliminary analysis of results, constructions of summary tables, and preparing first reports for the two panels. Identifying issues for further in-depth interviewing. Development of interview schedules. Conducting the interviews.

- Until end November: Integrated analysis of the quantitative and qualitative data from the surveys and the interviews. Preparation of a draft report to be discussed at the panel meeting half December.

- Until half of January, 2003: Writing the final draft. Submission to the ESF.

(3) *A Matrix of Evaluation Variables and Questions*. In the first section of this article it was argued that quality could best be conceived as a complex concept that needs to be further divided into distinct sub-dimensions. Here we present an example of such an analytical specification (the matrix is adapted from Arnold & Balazs, 1998)

¹ Here we like to express our appreciation to this external researcher, Dr Heide Hackmann, for her organisational and intellectual input in the design and execution of the review.

First we identified which *generic evaluation variables* had to be distinguished. The following units were identified:

- Policy, referring to the objectives, priorities and decisions (against which other evaluation units are examined);
- Resources, including financial resources, skilled personnel, expertise, capacities, instruments and infrastructure;
- Structure, referring to the organisation, composition of tasks and responsibilities, relationships, and division of labour;
- Process, the way the unit to be assessed operates, its functioning and management;
- Outputs, the direct results and products of this unit;
- Outcomes, referring to the effects of the outputs, the changes and benefits resulting from their availability.

Secondly the following *generic evaluation questions* were distinguished:

- Appropriateness, in which the question is asked: Is this suitable or the right thing to do? Appropriateness also includes the notion of adequacy;
- Effectiveness: Has it produced the expected results or effects?
- Efficiency: Does it function well or optimally, given the relation between input and output?
- Quality: How good or how satisfactory is it? (It will be noted that 'quality' is used here in a more restricted sense of an evaluation category, and not in the broad sense as implied in the title of this essay);
- Impact: What has happened as a result of it?
- Additionality: What has happened as its result over and above what would have happened anyway?
- Improvement: How can it be made better or strengthened, to which degree is there need for improvement, which changes have to be made?

The combination of these two sets of variables and questions leads to a matrix of *Generic Evaluation Questions applied to Units*, as visualised in Figure 1.

| Variables Related to | Questions about | | | | | | |
|----------------------|---------------------|-----------------------|--------------------|-------------|------------|-----------------------|-----------------|
| | Appropriate ness | Effec tive ness | Effi cien cy | Qua lity | Im pact | Addi tiona lity | Improve ment |
| Policy | | | | | | | |
| Resources | | | | | | | |
| Structure | | | | | | | |
| Process | | | | | | | |
| Outputs | | | | | | | |
| Outcomes | | | | | | | |

Figure 1. Generic evaluation questions applied to units (adapted from Arnold & Balazs, 1998).

This matrix has been used to guide the construction of the questionnaires and the interview schedules, and as a basic framework to evaluate the various units and responsibilities of the two ESF standing committees. As far as the latter are concerned, four separate categories to be evaluated were chosen: viz. the Standing Committee's Secretarial Unit, The Standing Committee in General, the Scientific activities, and Science Policy Activities, including the Forward Looks.

Of course, the 'units' in the generic matrix had to be specified for each of the four categories. For instance, the 'policy' unit for the Secretariat included budgetary planning, soliciting of applications, implementation of SC decisions, contact with Member Organizations, international networking, publicity, and policy development, whereas the 'policy' unit for the Standing Committee in General referred to its main tasks, such as budget accountability, maintaining links with other SC and/or ESF Units, contact with Member Organisations, international networking and publicity and marketing. Likewise, the 'resources' unit for the Scientific Activities is composed of its instruments, viz. exploratory workshops, networks, scientific programmes, EUROCORES, EURESCO conferences, and other (a.o., transatlantic collaboration), whereas this 'resources' unit for Science Policy Activities included the instrument Forward Looks, and the science policy skills and expertise of the Member Organisations.

Not all evaluation issues and all units are equally relevant and important for all four Standing Committee categories to be rated. In fact, an eventual scoring of all cells in the matrix for all four categories

would be both impossible and to a large extent useless. It was therefore decided that only the most relevant 'cells' of the matrix would be selected for each of the four categories. This was separately accomplished by the two panels for the Social Sciences and Humanities. Throughout the review process these 'purged' matrices have served as important guidelines for the focusing of the research attention to the relevant questions and issues.

Conclusion

In this article it was shown that 'quality' will remain one of the most salient criteria in reviewing educational systems and science research. At the same time it was argued that the concept of quality should be dissociated from the sphere of slogans and political catch-phrases, and that the term should be defined in logical and empirical terms. It was also shown that quality of more complex systems had to be assessed with the help of both quantitative indicators and qualitative, judgemental ratings. Important conditions for proper reviewing are expertise, independence and a strict scientific orientation. It was argued that Academies of Sciences and Humanities, and, at a European level, ALLEA as the European Federation of Academies, would be suitable institutions to conduct such quality assessment exercises. The review of two Standing Committees of the European Science Foundation was described as an example of such an endeavour.

References

- Arnold, E. & Balazs, K. (1998). *The evaluation of publicly funded basic research*. Brighton UK: Technopolis.
- Drenth, P.J.D. (1986). *Quality in higher education – evaluation and promotion*. CRE-Information, 18, 57-69.
- Groot, A.D. de (1983). Is de kwaliteit van hoger onderwijs te beoordelen? (Is it possible to evaluate the quality of higher education?). In: B.P.M. Creemers, et al. (Eds), *De Kwaliteit van het Hoger Onderwijs*. Groningen: Wolters Noordhof.