WISSENSCHAFTSRAT

Drs. 6285-04
Hamburg, 12 November 2004

Recommendations for rankings in the system of higher education and research

Part 1: Research

Inhalt — Seite

Introduction

The Federal Government and the state governments (*Länder*) commissioned the German Science Council in May 2003 to develop a concept for a ranking system. The Science Council established a working group in July 2003, which dealt with objectives of and methods for such a ranking system and held hearings on the issue with both national and international experts. The working group came to the conclusion that a proposal regarding methods for a comparative performance assessment must have clear goals and will only meet with acceptance if it is on a sound methodological footing. Therefore, the first step was for the group to develop recommendations for rankings in the system of higher education and research. From the point of view of science policy, procedures that meet these criteria are required both in teaching and in the research arena. As regards the field of teaching, preliminary work is needed to produce a definition for the term “quality” that can be used for comparison. Furthermore, the higher education sector is currently undergoing major transformations as a result of the Bologna process. In a first step towards specification, the working group has therefore prepared recommendations for a research rating system primarily designed for decision-makers at universities and non-university research institutions as well as for their partners at the various education ministries involved. This will be followed, in a second step, by a proposal for a procedure of comparative assessment in the field of teaching.

The working group included experts who are not members of the Science Council, for whose contributions we are particularly grateful.

The Science Council adopted these recommendations on 12 November 2004.

Summary

1. Comparative performance assessments by means of rankings and similar procedures may increase transparency with regard to performance in the system of higher education and research, inform the strategic decision-making of scientific institutions and provide a substantial input to effective and efficient competition.

Rankings help to document performance and current performance potentials. Combined with other instruments of strategic planning and quality assurance, they can inform the strategic decision-making of the various players in the system of higher education and research. However, managing the system of higher education and research solely or predominantly on the basis of ranking results is not recommendable, nor is this the aim of the procedure recommended by the Science Council.

Rankings in the narrow sense of the term, which involve ordinal rankings, only make sense if a certain set of specific conditions are fulfilled. In most cases, differentiation by means of ranking groups or according to a given grading scale (rating) is sufficient. The different interests of the various user groups, which are reflected in varying relative weights given to the various criteria, are taken into account through multidimensional assessments that lead to differentiated profiles for the rated institutions.

The recognised standard for comparative assessment procedures is a combination of peer review and quantitative indicators, the relative proportion of which may vary from case to case.

2. Given the importance of research activities for the success and international renown of scientific institutions, the Science Council recommends a research rating system for universities and non-university institutions that is capable of informing their strategic decision-making.

The research rating procedure should be subject-based. This requires a taxonomy that should take its cue from other taxonomies already in use at national or international level.

The assessment procedure should be multidimensional and be based on nine assessment criteria in the three dimensions of research, promotion of young research-

ers and knowledge transfer. The concept provides neither for a weighting of the research areas and criteria nor an aggregation of the results in an overall score.

The rating system should be based on a seven-point scale, in accordance with international standards, and be carried out by assessment panels for each individual research area. They also define the operationalisation of the criteria for each research area.

The rating is based on research area-specific profiles which need to be submitted by the rated institutions, as well as on bibliometric indicators. The assessment panels may define specific requirements with regard to the assessment data.

This procedure leads to research area-specific ratings based on the various criteria which allow universities and non-university research institutions to be compared with each other. At the same time, the results should also be such that they can be used to draw up performance profiles for the reviewed institutions.

The procedure would be supervised by a steering group that consists of renowned scientists and adequately represents the major scientific organisations. Organisational responsibility should lie with an organisation that has the required organisational and methodological competence in the field of research assessment and is independent of the rated institutions.

To test and refine the methods for the research rating system, the Science Council recommends carrying out a pilot study in two research areas. After successful conclusion of the study, the research rating system could successively cover all research areas on a rolling schedule of five to six years.

Furthermore, when the study has been concluded, Germany, along with other countries that have introduced similar procedures, should examine the question of whether it would be possible to implement a joint or combined rating system for research performance along the lines of international benchmarking. Such a system could help the countries involved to come to a more reliable assessment of the standing of their universities and non-university research institutions, controlling each other’s standards and learning from each other’s methods.

A. Documentation

A.I. Types and functions of comparative assessment procedures

In the last few decades, a comprehensive set of instruments for performance assessment and quality assurance has established itself in the German system of higher education and research. It includes first and foremost the various types of evaluation:

- The Max Planck Society has a system of quality assurance in place that consists of several stages and centres around periodic external evaluation of institutes by their advisory councils (*Fachbeiräte*). In addition, so-called "extended evaluations" take place at longer intervals covering several institutes dealing with similar research areas.[1]
- Since the 1980s, the Science Council had carried out systematic and periodic evaluations of the Blue List institutes in order to examine the prerequisites for joint funding by the Federal Government and the state governments. In 2003, responsibility for the institutions receiving joint funding was transferred to the Senate of the Leibniz Society. In future, each Leibniz institute will undergo evaluation by a working group established by the Senate committee on evaluation at intervals not exceeding seven years.[2]
- When the Helmholtz Association adopted a system of programme-based funding in 2001, it started reorganising its research activities in the form of research programmes which are subject to strategic evaluation at five-year intervals. In addition, the institutional evaluation procedures of the various research centres continue to exist.
- Between 1998 and 2001, the *Deutsche Forschungsgemeinschaft* and the funding organisations of the non-university research institutions underwent system evaluations focusing on their organisational structure and performance.[3]

[1] MPG (2002).
[2] Cf. www.wgl.de/evaluation.
[3] System evaluation of the Fraunhofer Society. Report of the Evaluation Commission, 1998; Promoting Research in Germany. Report of the International Commission for System Evaluation of the *Deutsche Forschungsgemeinschaft* and the Max Planck Society, Hanover 1999; Science Council: System Evaluation of the HGF – Statement by the Science Council on the *Helmholtz Gemeinschaft Deutscher Forschungszentren*, Cologne 2001; Science Council: Systemic Evaluation of the Blue List – Statement by the Science Council on the conclusion of the evaluation of Blue List institutions, Cologne 2001.

Many universities have established their own evaluation procedures, in some cases based on agreements with the states (Länder) in which they are located. In a number of states, evaluations of universities are centrally organised.[4]

By contrast, explicitly comparative assessments are far less widespread in Germany. They can be used as a supplement to the procedures related to individual institutions by pointing out their strengths and weaknesses and helping to put the results of internal assessments into perspective. Rankings that lead to ranking lists are common primarily in the field of university teaching. At present, the attention of the general public is above all on rankings of study courses at universities. They are published in news magazines and designed particularly for prospective students and their parents.

Rankings differ from evaluations in that they focus on the measuring and rating of outputs (rather than containing any recommendations for action or being process-oriented) and in that their purpose is to allow comparison, in other words, a number of institutions or funding programmes are analysed and assessed according to the same standards. Among the various procedures used for comparative performance assessments, the distinctive characteristics of rankings are a) a near-complete listing of the objects belonging to a given set (e.g. “universities in Germany”); b) operationalisation of performance criteria through a system of indicators; c) aggregation of the results of performance measurement by establishing ranking lists.5 The complete listing makes a difference between rankings and benchmarkings, in which the data of relevance for decision-making is obtained by means of comparison with selected reference institutions usually characterised by a particularly high level of performance. Rankings differ from ratings, which are an assessment of institutions on a predefined scale, usually carried out by expert groups, and which do not include the two aspects of operationalisation of indicators and aggregation in the form of ranking lists. Nevertheless, ratings can indeed lead to ranking groups or be included in rankings.

“Information aggregation” is a central function of rankings. The purpose of rankings is to provide information on a large number of heterogeneous institutions and assessment dimensions and present it in such a way that it may provide guidance in decision-making. Typically, rankings deal with institutions that compete with each other, and the decisions of the users of rankings are of fundamental importance for that

[4] For details on the procedure of the Scientific Commission of Lower Saxony, see p. 19.
[5] Bayer (1999).

competition. In other words, rankings are an instrument for making competitive systems more effective by increasing their transparency.

Transparency is a highly important aspect, not only for the players within the system, but also for users and funders. One of the most important user groups for higher education institutions are prospective students, who use rankings to choose their universities or, in other words, to find the ideal point of entry into the tertiary education sector – and thus, at least temporarily, the system of higher education and research. Future students are therefore the typical users of most existing national and international rankings – and simultaneously a "resource" for higher education institutions. Rankings also contain valuable information for foreign students or researchers, who are also "outsiders" to the German system of higher education and research, and it is conceivable that the availability of adequate information on the relative quality of the various institutions in a given country not only influences the choice of institution in that country, but indeed induces foreign students to opt to do part of their studies abroad in the first place. In this light, rankings can be regarded as an instrument for increasing the appeal of a country's system of higher education and research to international students.

As a rule, rankings only cover a certain section of the overall performance of the objects. Thus, when considering examples of international rankings and related procedures in the field of science, it is important to make a distinction between teaching rankings, which are usually designed for students, and research rankings.

A.II. International examples

In Germany, interest in rankings of scientific institutions has grown not least because rankings and related forms of comparative performance assessment have a long tradition and are taken very seriously in the Anglo-Saxon world. In the following, a brief description will be given of a few particularly well-known and at the same time highly representative rankings and similar procedures.

America's Best Colleges (U.S. News & World Report)

Together with a company named Common Data Set Initiative, U.S. News & World Report has published annual rankings of American universities and colleges since 1983. The results are sold in the form of a book or as a fee-based Internet service. The target group for these rankings are prospective students and students changing course who are trying to find the right education institution to improve their career prospects and quality of life.[6] Rankings thus boost competition between education institutions for students.

The U.S. News ranking reviews universities and colleges, ranking each institution as a whole without a breakdown according to subjects. However, in accordance with the Carnegie classification system[7], institutions are divided into four categories, for which separate rankings are drawn up: 1. national universities with the full range of degrees and a high level of research activity ("National Universities – Doctoral"); 2. national liberal arts colleges, with a focus on humanities and social sciences (at least 40% of degrees awarded) ("Liberal Arts Colleges – Bachelor's"), where students primarily seek Bachelor's degrees; 3. regional universities with a broad range of subjects and degrees, particularly Bachelor's degrees, to a lesser extent Master's and only rarely Doctor's degrees ("Universities – Master's"); and 4. comprehensive regional colleges ("Comprehensive Colleges – Bachelor's"), which are confined to Bachelor's courses in various subjects, but without the focus on humanities and social sciences typical of liberal arts colleges. In the case of the two regional categories, separate rankings are drawn up for North, South, Midwest and West, so that there are ten separate rankings altogether. In addition to the rankings, U.S. News & World Report processes data for arts and music colleges and universities and other specialised schools and publishes subject-based rankings of study programmes in the areas of engineering and business administration.

The rankings published by U.S. News & World Report are currently based on 16 quantitative indicators which are aggregated into seven categories, then weighted and finally added up to a composite weight score for the purpose of the rankings.[8]

---

[6] For more information on the Internet services of *U.S. News & World Report*, go to www.usnews.com.

[7] The purpose of this classification system, first published in 1973 and modified several times since (Carnegie Foundation 2001), is to divide higher education institutions in the US into relatively homogeneous types. It was originally designed for research on higher education, but has since generally established itself in the context of American higher education policy.

[8] Scores are adjusted by means of a system in which the best institution is given 100 points.

These seven categories and the weights assigned to each (given in per cent) are shown below:

- Academic reputation as established by a survey conducted among presidents, provosts and deans of higher education institutions: 25% [9]
- Faculty resources for teaching, based, among other things, on average salaries of teaching staff and student-staff ratios, as well as on student numbers in classrooms: 20%
- Percentage of freshmen continuing to the third semester (graduation and retention rate): 20%
- Student selectivity, assessed on the basis of entry exams and admission rates: 15%
- Financial resources: 10%
- Graduation rate performance: 5%
- Donations from alumni: 5%

The rankings are focused on the aspects of reputation, infrastructural prerequisites for teaching and student selectivity. In addition to performance-based rankings, which are drawn up using the weighting system described above, U.S. News & World Report also offers rankings based on cost-benefit ratio, which are again calculated by adding weighted scores and which specify the ratio of performance-based score to net costs for an average student (including tuition, accommodation, cost of living and scholarships), as well as for students receiving scholarships and students paying reduced tuition rates.

Research Doctorate Programs in the U.S. (National Research Council)

Since 1925, American scientists have repeatedly been asked to rate the quality of postgraduate studies at the various universities and colleges throughout the country. These surveys were originally conducted by a group of scientists and administrators and later adopted by research councils. In 1982, the National Research Council, an advisory body composed of members of the national academies, conducted a first study on the assessment of research doctorate programs in the US, commissioned

[9] This is a survey conducted among administrators. The term “peer assessment” used by the newspaper was chosen to make it clear that the institutions are rated by “peer institutions”, i.e. administrators of institutions belonging to the same category. This procedure should not be mixed up with the peer reviews customary among scientists, which would in any case be out of place in a non-subject-based ranking as that published by *U.S. News*.

by the Conference Board of Associated Research Councils; this study was repeated in modified form in 1993.[10] A new study is currently being prepared. A preliminary study on the methodology used for the assessment of doctoral research programmes has already been concluded.[11]

The studies carried out in 1982 and 1993 had three objectives and user groups:

- Supporting students and their advisers in looking for a suitable doctorate programme
- Providing decision-supporting information for university administrators, political decision-makers at federal and state level and managers of funding institutions
- Creating an up-to-date database for researchers dealing with the education system of the United States and its system of academic research.

Both studies centre around reputation data concerning the reviewed doctorate programmes, but there is one crucial difference: While, in 1982, the data were presented according to subject, with the names of the various universities and colleges listed in alphabetical order, the 1993 study for the first time showed rankings based on reputation scores. The rationale behind this choice of presentation was the claim that the alphabetical order previously used was “a source of frustration for many users”.[12]

The 1993 study comprises 41 subjects in the natural, engineering and social sciences and in the humanities. The subjects offered by so-called “professional schools”, for which a PhD is not the regular doctoral degree, i.e. law, business administration and medicine, were excluded, as were universities where these subjects produced less than five doctorates in five years. The study surveyed a total of 3634 doctorate programmes offered by 274 universities, thus altogether covering 90% of all doctoral degrees awarded in the said 41 subjects.

The academic reputation score was determined by surveying almost 17,000 researchers in the US. Each respondent was asked to rate a random sample of 50 doctoral programmes offered in his/her discipline according to the following three criteria:

- Academic quality of teaching staff (six-point scale)
- Effectiveness of training for scientists (4-point scale)

[10] Jones et al. (eds. 1982); Goldberger et al. (eds. 1995).
[11] Ostriker & Kuh (2003).
[12] Goldberger et al. (eds., 1995), p. 13.

– Qualitative change of the programme over the past five years (improved/unchanged/deteriorated)

Furthermore, respondents were asked to rate their degree of familiarity with the programme. The questionnaire included a list of the names of the academics participating in each programme and specified the number of doctoral degrees awarded under the programme over the past five years. The authors of the study sought to obtain a minimum of 100 ratings for each of the reviewed doctorate programmes. Because of the long experience with reputation ratings, the study includes a detailed discussion in which the authors advise caution in interpreting the results, drawing readers' attention to the fact that reputation ratings may be influenced by the sheer size of a programme, as well as to so-called "halo effects" and "superstar effects". [13]

This focus on academic reputation has not gone uncriticised in the US. Therefore, the 1993 study, rather than confining itself to mere reputation ratings, also contains a set of quantitative data on each doctorate programme:

Category	**Content**
Staff	Number (absolute), proportion of full-time professors proportion of researchers receiving third party funding proportion of award winners publications per faculty member, Gini Pub citations per faculty member, Gini Cite[14]
Students	Number (absolute), proportion of women number of doctorates
Graduates	Proportion of women, minorities, US citizens research assistants and teaching assistants average duration of doctoral studies (median)

There were considerations to record further statistical data, including on graduates' career success, but they were not put into practice.

The authors of the preliminary methodological study preparing the new study recommend carrying out periodic (annual) updates of an extended set of statistical data, leaving aside the specific question about the effectiveness of doctorate programmes

[13] A "halo effect" means that the good reputation of a superordinate institution benefits its sub-organisation. A "superstar effect" means an organisation benefitting from the good reputation of an outstanding individual faculty member or a group of such faculty members.

[14] The Gini coefficient is a concentration measure that shows whether absolute scores for a given indicator are based on the general performance of staff members or just the result of the work of a few outstanding scientists. Bibliometric data are only evaluated for the natural, social and engineering sciences (source: Institute of Scientific Information).

in the reputation survey[15], making greater efforts to measure the educational performance of doctorate programmes and choosing a form of presentation other than a mere ranking in order to avoid misinterpretation caused by a false impression of excessive accuracy (which was further reinforced by the showing of two decimal places for the median values in the academic reputation score). In order to compensate for the inert nature of reputation scores and the long intervals between surveys, the authors also analyse the dependency of the measured reputation on the more easily accessible quantitative data and propose using an equation developed by them to issue annual forecasts regarding changes in reputation during the period between two surveys.

Good University Guide (The Times)

Like the United States, the UK has a number of higher education rankings, published by newspapers and magazines and designed for prospective students. The most well-known and influential of them is the annually updated Times Good University Guide, which rates the study courses offered by British higher education institutions in more than 60 subjects. It also offers a global ranking of the 100 best universities in the United Kingdom.

The Good University Guide only uses a low number of indicators for the subject-based ranking lists:

- Score in the Teaching Quality Assessment (TQA)
- Score in the Research Assessment Exercise (RAE)
- Average school leaving certificate of first-year students (best three A level results)
- Percentage of graduates who find a job within the first six months of graduation or take up a postgraduate study course

The global list of the 100 best universities uses a number of additional indicators such as the student-staff ratio, library/computing spending, social and recreational activities, proportion of students awarded first and upper second degrees and overall graduation rates. All values are converted to a scale of 1-100, before the Teaching Quality Assessment (TQA) score is weighted with the factor 2.5, and the Research Assessment Exercise (RAE) score, with the factor 1.5.

[15] These responses show a high degree of correlation with the responses to the question about the scientific quality of faculty staff and, according to the Commission, are not based on any detailed knowledge of the structure and organisation of the doctorate programmes.

The remarkable thing about the Good University Guide is that – like the majority of comparable products offered in the UK, e.g. the Sunday Times University Guide or the Guardian University Guide –, it is largely based on an assessment of the quality of teaching and research through two procedures organised by the state: the Teaching Quality Assessment and the Research Assessment Exercise.

Teaching Quality Assessment (UK Quality Assurance Agency)

An important element of the reform of the higher education sector in Britain was the creation of Higher Education Funding Councils in England, Wales and Scotland in 1993, which, in a first step, established their own quality assurance procedures. The Quality Assurance Agency for Higher Education (QAA) was established in 1997 to create a uniform nationwide system of quality assurance, financed by contributions from higher education institutions. Since 2000, the QAA has continued the ratings of university study courses, previously known as Teaching Quality Assessment (TQA), under the name of "subject reviews", which are combined with institutional reviews or audits to add up to an "academic review".[16] Since the general public continues to use the term "Teaching Quality Assessment", it has been used here.

The primary objective of TQA is to ensure the quality of tertiary education and provide incentives for improvement, with the possibility of redistributing public funds on the basis of the results also being one of the options. TQA also aims at making information on the quality of tertiary education accessible to the public, thus meeting its obligation to render account to the public. Finally TQA is designed to help students find the right study courses.

TQA includes individual reports on the various study courses offered by the universities. In other words, these assessments are comprehensive evaluations rather than rankings in the narrow sense. However, these evaluations do comprise a rating on a predefined scale. In future, all study courses are to be assessed at intervals of six years. The results will then be summarised in a table that allows comparison of the assessments of the entire range of study courses offered in the higher education sector for a given subject.

The basis of each assessment is a self-assessment document submitted by the institution according to a predefined format, which contains information on the objectives

[16] Slight differences persist between the procedures applied in England, Wales and Scotland. Cf. Quality Assurance Agency for Higher Education (2000).

of the study course in question, a statement regarding the adequacy and clarity of these objectives, the effectiveness and quality of curricula, the quality of teaching and learning conditions and the aspect of quality assurance.

The study courses are assessed on the basis of these self-assessment documents by expert committees which are primarily composed of higher education teachers but may also include experts from industry and from industrial and other associations. The committees have to decide whether the institutions' objectives for the study course in question are adequate against the so-called "subject benchmarks" and whether institutions achieve their own objectives. Study courses are rated along six dimensions:

- Curriculum design, content and organisation
- Teaching, learning and assessment
- Student progression and achievement
- Student support and guidance
- Learning resources
- Quality assurance and enhancement

In each dimension, assessments follow a four-point scale, which provides information as to whether the institution's efforts contribute fully, substantially or partly to the achievement of its objectives or do not do so at all. Members of the expert committee are required to seek all information they may need to substantiate their judgements; this may include site visits, participation in advisory board meetings of the reviewed institution, interviews with students, consideration of results from other (internal) assessments or the studying of tests, final exams or course materials. This task is to be accomplished at a relatively low expense where the outcome of the overall assessment is clear, and the cost should only be allowed to be higher in more complicated cases. As a rule, quantitative data is collected, too, but there is no predefined set of indicators. The assessment reports lead to a final statement, in which study courses are rated as "commendable", "approved" or "failing".[17] There is no comparison with the study courses offered by other universities.

[17] In Scotland, the ratings are "excellent", "highly satisfactory" and "satisfactory". The website of the Scottish Funding Council does not mention a category for unsatisfactory study courses.

Research Assessment Exercise (UK Funding Bodies)

Since 1986, research activities at higher education institutions in the UK have been assessed at five-year intervals in order to inform decision-making on the distribution of basic funding for research. The Research Assessment Exercise (RAE) is thus part of the dual support system in which research funding comes in the form of basic grants provided by the Funding Councils, on the one hand, and project funding provided by the Research Councils, on the other. The basic funding for teaching also comes from the Funding Councils, but primarily on the basis of aspects related to capacity rather than the results of the TQA (see above).

In addition to the selective distribution of basic research funding, one of the secondary objectives of the RAE is to provide different user groups at higher education institutions and among the general public with information on the quality of research in the British higher education sector.

Assessment of research activities at higher education institutions takes place in 68 discipline-based units of assessment.[18] It is for each university to decide in which of these units it makes a submission and which of its faculty members it reports as an active researcher in which unit. Research funding is only allocated to those faculty members who are registered in a submission. However, there is no obligation to register all faculty members for the purpose of the RAE. The components of a submission of research activities in a unit of assessment are shown in the following table:

Category	**Content**
Staff information	summaries of all academic staff details of research-active staff research support staff and research assistants
Research outputs	up to four items of research output for each researcher
Description	information about the research environment structure and policies strategies for research development qualitative information on research performance and measures of esteem
Data	amounts and sources of research funding numbers of research students number and sources of research studentships number of research degrees awarded indicators of peer esteem

[18] In the following, all figures will refer to the latest RAE, the results of which were published in 2001 (RAE 2001).

For each unit of assessment[19], a panel is established consisting of 9 to 18 scientists, predominantly from the academic community, but also comprising experts representing the private sector. In order to secure the assessment of international research excellence, a number of foreign experts are consulted. The panels rate all higher education institutions that have reported activities in their unit of assessment on the basis of that submission, taking due account of the research outputs submitted. The are no visits to the institutions.

Submissions are rated on a scale of 1 to 5*, with the individual grades being defined by the proportion of research activities meeting national or international standards of excellence. The panels agree in advance on the weighting of the various types of data and on a set of assessment principles. This decision is published, in some cases along with a statement on issues that should be particularly taken into account in the submissions. The purpose of this procedure is to reflect the different criteria of assessment in the various subjects.

No ranking is established – indeed the RAE is not a ranking, but a rating system. However, unlike in the case of TQA, all universities and all subjects are rated simultaneously, so that the grades published in the form of tables can be read as ranking groups.

The grades awarded are converted into funding factors which are incorporated into a formula for the provision of basic research funds by the Funding Councils.[20] The basis of assessment is the number of faculty members reported as active researchers, and a distinction is made between cost-effective and cost-intensive units of assessment by using a subject-specific factor between 1 and 1.6.

The RAE concept was fundamentally reviewed between 2002 and 2004 (HEFCE 2003). One of the results was the adoption of a new system of assessment (RAE 2004). Under the new regime, a university's research activities in a given unit of assessment will no longer receive an overall rating on a seven-point scale; instead, the review will show the proportion of overall research activity reported in a submission that meets each of four defined levels of quality (one, two, three and four star). The purpose of the rating system is no longer to represent this so-called research profile

[19] With the exception of a few joint panels; responsibility for the 68 units of assessment lies with 60 sub-panels.
[20] For research activities receiving RAE grade 3a, the funding factor in 2003 was 0, for grade 4 = 1 for grade 5 = 2.793 and for grade 5* = 3.357.

in the form of a one-dimensional diagram. Instead, it will be up to the users whether they are interested in the median value of research quality, the proportion of cutting edge research or the institution's total capacity.

Another modification recommended was to give panels greater autonomy in defining their criteria. The panels will be expected to make a particular effort to develop quantitative indicators that can be calculated on the basis of the standard data to be submitted or with the help of existing databases, and to develop criteria for assessing practice-based and applied research. Likewise, the data and materials to be submitted can be modified by each panel according to the criteria it has set itself. Thus it will in future be possible to reduce the number of research outputs reported per faculty member compared to the present system (e.g. two or three instead of four). Furthermore, there will be a new option to submit a number of research outputs as teamwork produced by a group of researchers.

The RAE review showed that it is difficult to maintain uniform standards of assessment across the different research disciplines. In order to improve the consistency of standards, a new category of 15 to 20 main panels will be created to complement the work of the discipline-based sub-panels, of which there will be approx. 70. This new category of panels will both control the criteria of the sub-panels and make the final decision on the ratings of the various research activities.

The review confirmed the direct link to the allocation of basic research funds as a primordial goal of the RAE. Where possible, a greater degree of transparency in the allocation process is to be created by a preliminary statement of the Funding Councils on the funding factors.

Netherlands Standard Evaluation Protocol (NWO, VSNU, KNAW)

In the Netherlands, all research activities receiving public funding will be assessed at six-year intervals[21] as of 2003, following a joint initiative launched by the Association of Dutch Universities (VSNU), the Netherlands Organisation for Scientific Research (NWO) and the Royal Netherlands Academy of Arts and Sciences (KNAW). The initiative has a triple objective: to improve the quality of research; to improve the quality of research management; and to ensure the accountability of research institutions towards their funding organisations, sponsors and Dutch society at large[22]. With re-

[21] In parallel, all universities will carry out their own quality assessments with regard to teaching.
[22] NWO, VSNU, KNAW (2002).

gard to the first two objectives, the players in this process are the funding organisations, the heads of the various institutes and the responsible researchers.

The assessment applies to institutes of universities and non-university institutions in the sphere of competence of NWO and KNAW which, due to different organisational structures of research in these sectors, are only defined in very general terms as “groups of researchers with a shared mission operating under a common management”. The activities of each institute are further subdivided into thematically coherent research programmes.

Each organisation, i.e. NWO, KNAW and the various universities, is responsible for assessing the institutions within its sphere of competence in accordance with the agreed standards. An institute is rated by an assessment group the composition of which depends on the institute’s overall mission.

The basis of assessment is a self-evaluation document submitted by each institute, which contains standardised information about both the faculty as a whole and its various research programmes:

Level of aggregation	**Information**
Institute	Overall mission of the institute Organisational structure and management Strategy and tactics Staff Resources, funding, infrastructure Processes of research, internal and external cooperation Academic reputation Internal evaluation External validation of self-evaluation
Research programme	Organisational structure and management Strategy and tactics Processes of research, internal and external cooperation Academic reputation Internal evaluation Staff Resources, funding, infrastructure Three to five publications or other research outputs showing the quality of the research Complete list of publications Quantitative overview of publications by category

In addition, each institute is requested to submit a self-assessment along the lines of a SWOT analysis (strengths, weaknesses, opportunities and threats) and build a strategy on that basis.

After receiving the self-evaluation document, the assessment group visits the institute and meets its director or board, the heads of the various departments (research programmes), the advisory council and any other persons or groups requesting a meeting with the reviewers.

The assessment leads to a written assessment report that has a prescribed format and contains assessments of both the institute as a whole and the various research programmes. The latter are rated on a five-point scale on the basis of the criteria of quality, productivity, relevance and vitality. The grades on that scale are given a verbal definition, with “excellent” signifying an internationally leading role, and “very good”, an internationally competitive, nationally leading activity. The four criteria are further subdivided into sub-criteria to ensure a comprehensive assessment of all aspects relevant to the research activity.

The assessment report, along with the self-evaluation document, is sent to the governing board of the funding organisations (NWO, KNAW) or to the university council, which will then draw the necessary conclusions for the institute’s future.

The assessment report, the self-evaluation document and the statement of the governing board together constitute the result of the research assessment process. They are to be published as early as possible.

The use of a standardised scale allows a comparison to be made between the various institutes and research programmes in a given research area. Compliance with uniform standards is to be verified by means of a meta-evaluation carried out by an independent commission.

A.III. Rankings in the German system of higher education and research

The history of rankings in the German system of higher education and research began in the 1970s. During the initial stage, the focus of attention was on rankings of entire universities. A broad range of methods were tested in various subjects and disciplines over the years. Discipline-based rankings covering a broad range of subjects have been carried out at regular intervals for the last 15 years or so. The attention of the general public was primarily attracted by the rankings published by the large news magazines, beginning with the ranking published by Der Spiegel in

1989.[23] The most prominent example at present is the higher education institutions ranking drawn up annually by the Centrum für Hochschulentwicklung (CHE) and published as a university guide in the magazine Der Stern.

Like most of the international examples described above, the primary target group of the rankings published by German news magazines are prospective students and their parents who are seeking relevant information to guide their choice of university. In nationwide surveys carried out several years ago, more than a quarter of first-year students said that their university's good academic reputation, documented in rankings, had played a major role in their decision. An evaluation of data obtained from the Central Admissions Office (ZVS) showed that, as a result of the publication of such rankings, the number of applicants to universities with better ratings increased by 20%.[24] Nevertheless, the most important factors determining the choice of university – at least in the case of students doing economics or business studies, for which such data is available – are still the attractiveness of the university town, and especially proximity to home, and the local cost of living.[25]

In recent years, the research activities of universities have increasingly become the focus of rankings. In 1997 and 2000, the Deutsche Forschungsgemeinschaft published differentiated reports on the projects receiving its funds, thus building a bridge to the rankings of the news magazines. DFG extended this system of reporting in 2003 by publishing an updated version providing further data in the form of a funding ranking, which contains basic data on research activities receiving public funding.[26] Based on its ranking of HEIs, CHE published a Research Ranking in 2002 which is primarily designed for scientific researchers including young scientists. The update of this Research Ranking, published in 2003, also sought to identify the best research universities in Germany.[27]

Earlier rankings of entire universities

According to Daniel (1998), earlier examples of research rankings of entire universities in Germany include

[23] According to Rosigkeit (1997), the earliest example is a ranking published by the Austrian magazine *Der Wiener – Zeitschrift für Zeitgeist* in 1987.
[24] Daniel (2001).
[25] Büttner et al. (2002), Fabel et al. (2002).
[26] Deutsche Forschungsgemeinschaft (2003), p. 5.
[27] Centrum für Hochschulentwicklung (2002), Berghoff et al. (2003b).

- a comparison of German universities based on numbers of publications and a citation index relying on data from the Science Citation Index and carried out by Spiegel-Rösing on behalf of the Federal Ministry of Education and Science in 1975;
- the university rankings of the Research at the Universities survey, based on professors' publication productivity, which is assessed on the basis of average publication levels in each discipline, and established on the basis of a representative survey by the Allensbach Institute in the 1976-77 winter semester;
- the popularity rating of German universities established and published by the Alexander von Humboldt Foundation after evaluating their admission notices in 1981.

The common characteristic of the three initiatives is that each of them compares entire universities and uses only a small set of indicators. Given the low level of coincidence between the published rankings (Daniel l.c.), the irregular intervals between the initiatives and their open objectives, this period marks the early experimental phase of the ranking methodology in the German system of higher education and research.

The Science Council discussed the issue of introducing ranking procedures to increase the transparency of the system of higher education and research as a prerequisite for competition[28] – without, however, producing tangible results at the time.

The CHE University Ranking

The purpose of the University Ranking, first drawn up in 1998 and published annually since 1999 by the *Centrum für Hochschulentwicklung* as a university guide in the magazine *Der Stern*, is to guide prospective students in choosing their future universities and increase the transparency of higher education institutions with regard to courses and performance.[29] The CHE higher education institutions ranking is based

[28] In order to increase the transparency of the system of higher education and research, the Science Council recommended two steps in 1985: firstly, self-portraits of the universities at regular intervals and secondly, a comparative assessment of performance. The report reads: "The assessment procedure could take its cue from ranking methods developed by the American higher education system," adding that "the system can only reasonably be applied to individual subjects, not universities as a whole. Furthermore, it is important to avoid overstating the importance of individual indicators and base ratings on a broad spectrum of different indicators. Moreover it does not seem necessary to produce a ranking in which each faculty has its rank; instead, ranking groups should suffice". (Science Council 1985, p. 27).

[29] Stern (2003); Berghoff et al. (2003a).

on a multidimensional decision-making model and comprises subject-specific data on study conditions, a few research indicators and ratings from both professors and students, without, however, aggregating the data in an overall ranking list.[30] CHE believes that an overall rating given to each university would end up blurring any differentiated assessment of research, teaching, learning support, resources, etc. Even with regard to individual indicators, CHE does not produce rankings in the strict sense of the term, but instead presents the results in the form of group rankings (top, medium and bottom group). However, on the basis of a selected set of criteria, a small group of universities are marked as recommended ("study tips") for three different types of students: single-minded, research and practical students.

The units of assessment are study courses at individual universities, and the underlying data and survey results refer partly to areas of study, partly to faculties and to study courses. The studies guide published in 2003 includes CHE rankings of different dates for 34 subject areas.

CHE presents the various criteria of the underlying decision-making model in a matrix:

Students	**Study results**	**International orientation**
Research performance	Teaching and learning	Resources
Occupational relevance, labour market	Overall verdict by students, professors' recommendations	Information on university and university town

Each of these nine dimensions has several indicators assigned to it:

1. Students: This dimension is characterised by information on the number of students, first-semester students and applicants and the proportion of women students and dropout rates.

2. Study results: This dimension provides information on absolute numbers of graduates, average grades, average duration of studies (median) and classification of graduates according to the number of semesters studied.

3. International orientation: Information in this dimension includes data on the possibility of obtaining double degrees, on participation in the European Credit Transfer

[30] For a scientific assessment of the methods used for the higher education institutions ranking, see Hornbostel (2001).

System, on the existence of subject-specific foreign language classes, on mandatory semesters abroad and exchange programmes as well as on the proportion of foreign students and visiting professors.

4. Research performance: This dimension includes data and assessments. The data show third party funding per researcher/professor, for some subjects patents per professor, publications per professor as well as doctorates and (sometimes) habilitations per professor. In addition, the results of a survey among professors are presented, in which professors in each subject were asked to name up to three higher education institutions that they believe occupy a leading position in that subject in Germany.

5. Teaching and learning: As well as data on student-staff ratio (students per professor), implementation of evaluations regarding teaching quality and a number of other services relevant to study success, this dimension includes the results of student questionnaires concerning several aspects of study conditions: study guidance, study courses, support from staff, communication between students and staff as well as among students and content of curricula and examinations.

6. Resources: This dimension, too, offers data and ratings from students. The data is subject-specific and includes details of IT infrastructure spending, provision of laboratories, non-financial resources, number of beds, etc. The questionnaires given to students deal with the quality and availability of PCs, workstations and laboratory facilities, the quality of libraries, the quality of available premises and the availability of audiovisual media.

7. Occupational relevance and labour market: This dimension provides aggregated data on specific courses with occupational and labour market relevance (number of weekly lessons per 100 students) as well as students' overall assessment of the measures to promote occupational and labour market relevance.

8. Overall verdict: This dimension contains the responses of students to the question regarding the overall situation in their subjects as well as the so-called "professor's recommendations", where professors are asked to name up to three higher education institutions worth recommending in their subjects.

9. University and university town: In addition to the subject-specific information, CHE compiles a set of data on population size, proportion of students, their housing

situation and public transport infrastructure in the university town, as well as on student numbers, semester tuition and general services provided by the university (study guidance, sports activities, etc.).

The printed version of the CHE University Ranking only provides a limited number of indicators per subject, which are considered particularly relevant. By contrast, users of the Internet version can compile individual rankings based on up to five indicators selected from the entire range of data provided.

The CHE Research Ranking

Since 2002, CHE has also published Research Rankings on the basis of a selective evaluation of data obtained in the course of its University Rankings. Their purpose as stated by CHE is to "identify the research-active faculties in Germany."[31]

The 2003 CHE Research Ranking[32] covers 13 subjects in the natural and social sciences and the humanities. Although rankings have been established for mechanical and electrical engineering, they are not part of the current publication, for various methodological reasons.

The CHE Research Ranking includes three indicators for research activity: volume of third party funding, number of publications and number of doctorates per professor. In addition, citations are recorded for four subjects (biology, chemistry, physics, pharmacy). The data on the volume of third party funding refer to the funds used by a faculty in a given subject over a period of three years as determined by means of a survey. The number of publications is determined through a bibliometric analysis with the help of relevant databases; for certain subjects, publications were weighted by type and length.[33] Numbers of doctorates were determined for a period of four semesters by means of faculty surveys. All indicators in the CHE Research Ranking are quoted in absolute numbers and in proportion to the number of researchers (third party funding) or professors (publications, doctorates).

Reputations were determined by means of a survey among professors in which respondents were asked to name three universities which they would recommend for studies in their subjects or which they consider to occupy a leading position in re-

[31] For details go to www.dashochschulranking.de/allgemeines_fr.php, 15 December 2003.
[32] Berghoff et al. (2003b).
[33] For details on criticism of the publication indicators used by CHE, see Ursprung (2003); for a contrary opinion, see Berghoff & Hornbostel (2003).

search in these subjects. In its Research Ranking, CHE classifies universities that are named by more than 5 per cent of respondents as having a good reputation.

CHE defines faculties with a strong research performance by establishing sub-rankings for each of the indicators and identifying top groups from these lists. In the case of the absolute indicators, the top group is the group of those universities in the highest ranks whose combined scores add up to at least 50 per cent of the sum total for each indicator. In the case of the relative indicators, the top group consists of the universities in the first quartile of the ranking. Faculties are credited with a strong research performance if they are ranked in the top group with regard to at least half of the indicators used for the subject in question (absolute and relative indicators, but not counting reputation).

In a further aggregation step, CHE defines research universities in the humanities and natural sciences as meaning those universities of which at least 50 per cent of the faculties registered in the 13 CHE rankings are credited with a strong research performance. Based on this criterion, CHE has identified seven universities in Germany as research-active universities in 2003.

The DFG Funding Ranking

In autumn 1996, the Deutsche Forschungsgemeinschaft, following an initiative by a group of university presidents, published data on the ten higher education institutions that received the greatest volume of funding from DFG in the period 1991-1995. This publication prompted a lively debate, which soon caused DFG, in consultation with the Rectors' Conference, to draw up a broader report on its funding policy, thus responding to a high level of interest in comparative data. In the resulting publication[34], express reference is made to the suitability of data on third party funding as an indicator of research activity or – where, as in the case of DFG, applications are processed by a review group – of research performance. Although DFG never claimed to present a research ranking of HE institutions, the publication was received as such by the general public, not least because of its presentation, in which HE institutions were listed in the annexed table not in alphabetical order, but in descending order based on the volume of third party funding received in absolute terms or per professor.

[34] *Deutsche Forschungsgemeinschaft* (1997).

In 2000, a follow-up report was published with a considerably broader scope including non-university institutions, followed in 2003 by a publication entitled “Funding Ranking”, which, for the first time, included not only data provided by DFG itself and the Federal Statistical Office, but also data from the Alexander von Humboldt Foundation, the German Academic Exchange Service (DAAD) and the European liaison office of the German research organisations (KoWi), as well as data from bibliometric analyses.[35]

Responding to the spread of rankings in Germany, the intention behind DFG’s publication is to place the debate on the assessment of research on a broader footing, thus contributing to the definition of “best practice” for the establishment of rankings of academic excellence. The results of the study are summarised under five headings:

1. At the centre of the Funding Ranking is an analysis of DFG funding approved for universities and non-university institutions between 1999 and 2001. As well as absolute amounts approved per institution, expressed as sum totals and broken down according to subject and research area, data for the individual subjects include data on funds approved for each university and programme group[36] and on funds approved per university and professor, or per university and researcher. The ranking also looks into the ratio of funding provided by DFG to the total third party funding received by the universities. The results show a strong correlation at the level of the institutions, whereas there are significant differences between the various research areas. In the case of non-university institutions, the funding approvals are shown for each scientific field and for each programme group.

2. As a new addition in 2003, an analysis was carried out of networked research on the basis of data on the provision of funding under coordinated programmes of DFG.[37] The object of evaluation is the joint participation of institutions in coordinated programmes as established on the basis of the institutional addresses of sub-project leaders. The centrality of institutions within academic networks in the various fields of science is determined by the number of its partner institutions. By

[35] *Deutsche Forschungsgemeinschaft* (2003).
[36] DFG has grouped its funding instruments according to structural criteria in the following programme groups: “Individual Grants Programme”, “Coordinated Programmes”, “Direct Support for Young Researchers” and “Scientific Prizes & Awards”.
[37] These include special research areas, priority programmes, research groups, postgraduate studies and humanities research centres. At the time the data were recorded, the DFG Research Centres programme had not been established yet.

showing links between institutions jointly participating in a certain minimum number of coordinated projects (that varies from field to field), the report illustrates the core structures of cooperation networks in DFG projects. Furthermore the online version of the Funding Ranking also shows a dynamic representation which allows the cooperation relations between individual institutions to be traced.

3. Another new element is an analysis of the institutions employing the various DFG reviewers, which DFG believes is a very good indicator of the research expertise concentrated in one place. The results are based on the written reviews requested by DFG for the relevant period. The Funding Ranking shows the absolute number of reviewers for each university and scientific discipline as well as for each university and research area. In addition, the correlation between number of reviewers and amounts granted is analysed.

4. DFG provides a whole range of data on the international orientation of the research activities: the number of the visiting researchers for each university and scientific discipline as well as for each university and research area (1997-2001); the number of visiting researchers receiving funding in relation to the number of professors or researchers employed by the university in each scientific discipline; analogue data for the German Academic Exchange Service (2000-2001); and data on the participation of German universities in the Fifth RTD Framework Programme of the EU (1998-2002). In addition, DFG provides data on the nationality of visiting researchers and the nationality of cooperation partners in the EU projects, without, however, giving a breakdown for the various institutions within Germany.

5. The fifth heading in the Funding Ranking concerns bibliometric findings. Here DFG has evaluated two published studies[38], but confined itself to analysing the correlation between DFG approvals and publications (citations, if available, counted at the aggregation level of the university as a whole and for universities in the field of medicine). Furthermore, it has analysed the number of publications per professor/researcher and the relative citation index in the field of medicine, however only for DFG approval ranking groups, but not for individual universities.

[38] DFG used a study published by the Swiss Centre for Science and Technology Studies (CEST) on the number of publications from universities in general and a study by the Netherlands Centre for Science and Technology Studies (CWTS) on publications and citations in medical research.

In its summary, DFG has presented the results of the various analyses in the form of a comparison between ranking groups of universities for the most important indicators analysed (in absolute numbers and apportioned to the various professors) and is thus also offering a multidimensional ranking.

Research evaluation at Lower Saxon universities and non-university institutions

Since 1999, the Scientific Commission of Lower Saxony has carried out a state-wide evaluation procedure based on uniform criteria.[39] The purpose of that procedure is to

- support universities in developing their own research profiles;
- provide universities with criteria for the planning and implementation of quality assurance and improvement measures;
- improve the contribution of state governments to the profile-building of universities and draw up guidelines for the structural planning at universities;
- contribute to the development of criteria for quality-oriented funding at the state level;
- increase transparency with regard to universities' performance.

The research evaluations are subject-based.[40] A review group is established for each subject to be evaluated. The evaluations are carried out on the basis of standardised self-reports submitted by the universities and of visits by the review groups involving meetings with university governing boards, faculties, subject representatives and individual research units or researchers. In this regard, it is a procedure of the Informed Peer Review type.

Reviewers are required to assess institutions' research performance in terms of quality and relevance as well as effectiveness and efficiency. The criteria of quality and relevance are subdivided into the following sub-criteria:

- Innovativeness of research
- Scientific impact
- Interdisciplinarity
- Cooperation arrangements with other research institutions

[39] Scientific Commission of Lower Saxony (2002).
[40] Non-university research institutions financed by the state of Lower Saxony are evaluated as entire institutions.

- International cooperation
- Promotion of young researchers
- Cooperation with industry, public authorities and cultural institutions

In the evaluations, the research quality of the various research units is rated on a scale from 1 to 10, with 10 being the best score for research performance. There is no overall rating for entire universities. All rating information is primarily designed to ensure uniform standards for the evaluation of the various universities.

Each review group's evaluation results and recommendations are summarised in a final report, which is transmitted to the university for comments. The steering group in charge of the procedure discusses the final reports and the comments submitted by universities and forwards them, if necessary, with additional comments, to the Scientific Commission of Lower Saxony. The commission discusses evaluation results from a structural perspective and issues recommendations to the state government and, in some cases, to individual universities.

University ranking in terms of gender equality

In 2003, the Centre of Excellence Women in Science published a university ranking on the basis of gender equality aspects (CEWS 2003). Taking its cue from the gender equality requirement laid down in the Higher Education Framework Act, the ranking is designed to provide universities with a comparative yardstick that enables them to rate their own performance in the area of gender equality and equal opportunities. The primary target group of the ranking are university governing boards and managers, but also federal and state ministries, scientific organisations and policymakers that have an interest in individual universities meeting gender targets.

The university ranking based on gender equality aspects rates universities, technical colleges and art academies as entire institutions, but each within its separate category. It uses only quantitative indicators of a certain type, i.e. the proportion of women and its development over time in the various status groups. Thus the proportion of women is recorded in the following groups:

- Students
- Doctoral students
- Habilitations
- Full-time employment scientific and artistic staff

- Professors
- Changes in full-time employment scientific and artistic staff over time
- Changes with regard to professors over time

As in the case of the CHE ranking, rankings are established for each indicator, and top (first quartile), medium (second and third quartiles) and bottom (fourth quartile) groups are identified. A simple points system (2 points for each rating in the top group, 1 point for each rating in the medium group) is used to add up all seven indicators to produce a composite score for each university. As a result, universities, technical colleges and arts colleges are ranked in groups.

A.IV. Comparison of existing comparative assessment procedures

A tabled comparison between various ranking and rating procedures for the fields of teaching and research shows that they are characterised by a host of different combinations of aims and features. The spectrum ranges from ranking procedures that strictly conform to the definition laid down earlier – e.g. the *U.S. News* ranking, which is solely based on indicators and leads to the establishment of ordinal rankings – to the peer-review-based procedure in the Netherlands, which, despite being uniform nationwide and allowing comparison due to its standardised point scale, is closer to an evaluation procedure in other respects, such as self-evaluation, mandatory visits to the institution and presentation of results in the form of reports on individual institutions.

There would be even more possible combinations of elements of procedures if certain other procedures were included. Thus, for example, the combination “research – entire universities – quantitative indicators – ranking”, which is not listed in the table, has been implemented several times, for example in the ranking of 500 international universities carried out and published on the Internet by academics from the Shanghai Jiao Tong University in 2003 and updated in 2004, and in the Champions League of Research Institutions published by the Swiss Centre for Science and Technology Studies (CEST). An international comparison alone does not allow any conclusions to be drawn about which procedural elements are indispensable for a given dimension of performance or set of objectives. Nevertheless, three trends are worth noting.

Firstly, rankings in a narrow sense are becoming less and less common. Given the uncertainty of the data for certain indicators and the problem of weighting between

them, presentation of data in rankings is considered inappropriate, since the implied level of accuracy cannot be delivered. An alternative consists in presenting ranking groups in order to prevent the importance of ordinal lists being overstated. The same can be achieved by stating confidence intervals for ranks in a list. Such ranking groups are based either on percentiles, usually designating the top and bottom quartiles of the distribution, or on confidence intervals, which ensure a statistically relevant difference between top group and bottom group. Such ranking groups based on confidence intervals were used for the SPIEGEL and CHE rankings. Ranking groups were also proposed when the methodology for assessing doctoral programmes was revised by the National Research Council (Ostriker & Kuh (eds. 2003)). Another possibility is a rating on a predefined grading scale – the obvious solution for peer-review-based procedures.

The second trend militating against an aggregation of information in the form of ordinal rankings is a growing tendency to make ratings more transparent for users, rather than adding up various ratings based on different criteria to a composite score. This can be achieved through a multidimensional representation that allows the position of the assessed units to be illustrated along several different dimensions of assessment.

A third trend is the increasing tendency to use expert assessments along with quantitative indicators. Peer reviews have become an international standard in assessing the quality of scientific research. However, there are limits to their reliability.41 In any case, reviewers need a solid basis of data to prevent the results of ratings becoming mere assessments of reputation, with all their drawbacks (subjectivity, delay, halo and superstar effects).[42]

[41] Bornmann & Daniel (2003).
[42] Cf. section “Research Doctorate Programmes in the US”.

	U.S. News	Research Doctorate Programs 1993 (U.S. NRC)	Research Assessment Exercise 2001 (UK)	Netherlands Standard Evaluation Protocol 2003-2009 (NWO, VSNU, KNAW)	CHE University Ranking	CHE Research Ranking	DFG Funding Ranking
Performance dimension	Teaching	Promotion of young scientists, research	Research	Research	Teaching	Research	Research
Objectives	Student guidance	Strategic instrument; basis for educational research; university selection guidance	Efficient distribution of basic funding for research	Strategic instrument; meet the obligation of research organisations to render account to the public	Student guidance	Identify research-active faculties and universities	Boost competition between research institutions; contribute to development of methods for research rankings; meet DFG's obligation to render account to the public
Objects	Universities and colleges	Doctoral programmes of universities by subject	Research activities of universities reviewed by subject	Institutes, research programmes of universities and non-university research institutions	Faculties, teaching units, study courses reviewed by subject	Faculties by subject	Research activities of universities and non-university institutions, by subject
Basis of assessment	Quantitative indicators	Academic reputation, quantitative indicators	Informed peer review	Informed peer review, site visits	Quantitative and qualitative indicators, academic reputation	Quantitative indicators, academic reputation	Quantitative indicators
Results	Rankings (by category of institution)	Rankings (by subject)	Rankings, ranking groups (by subject)	Assessment reports, rating (based on performance criteria)	Multidimensional ranking groups (by subject and other criteria)	Multidimensional ranking groups (by subject and indicator); identification of research-active faculties and universities	Multidimensional rankings (by research area and indicator); network analyses; multidimensional ranking groups

In the case of comparative assessments on a nationwide or international scale in particular, quantitative indicators are indispensable to limit demands on the time of reviewers and institutions and obtain reliable results despite time constraints. In the area of research, publication and citation indexes have established themselves as criteria, along with data on third party funding.[43]

However, assessments solely based on indicators are regularly criticised, because in most cases, the indicators themselves require competent interpretation. Therefore, in spite of their benefits (scales, low expense, objectivity), quantitative indicators are rarely used as the sole basis for rankings. If at all, this is most likely to happen where a high level of differentiation coincides with a low level of market transparency (U.S. News), where the predominant objective is to provide general information and increase transparency (CHE Research Ranking) or where a comparison is to be made at a very high level of aggregation (Shanghai Jian Tong University ranking in 2003). By contrast, in procedures explicitly designed to inform strategic processes and allocation decisions (RAE, Netherlands Standard Evaluation Protocol), quantitative indicators tend to be used to support the assessment of experts (informed peer review).

[43] Van Raan (1996), Hornbostel (1997); for an example of application in practice, see Tijssen, van Leeuwen & van Raan (2002).

B. Recommendations

B.I. Preliminary remarks

The availability of comparative quality information on the various services provided by universities and non-university research institutions plays a crucial role in ensuring effective and efficient competition in the academic field. Since such quality assessments require highly specialised knowledge, systematic procedures are needed to provide decision-makers with the relevant information. Comparative performance assessments are therefore an essential component of a reform process that reinforces the autonomy of academic institutions and involves a transformation from detailed state control to a global system comprising elements of competition.

In the medium term, comparative assessments should cover all relevant dimensions of the performance of universities and non-university research institutions, taking account of the various user groups using the assessment results. However, for methodological reasons – including the fact that subject classification in research does not necessarily coincide with subject classification in teaching, which requires different object definitions –, a simultaneous assessment of all types of performance, in particular in research and teaching, under a uniform procedure is not recommendable.

The Science Council believes that at present, more reliable and detailed comparative data on the quality of performance is needed in the field of research in order to make competition between universities as well as between universities and non-university institutions more transparent and allow a lasting improvement in overall performance. It therefore recommends establishing a research rating system in Germany (B.III). This procedure, however, should also assess the quality of measures to promote young researchers and knowledge transfer, as these aspects are closely related to overall research performance.

For reasons of quality assurance, the existence of alternative ranking and rating initiatives should be welcomed. The level of acceptance among users as well as among rated institutions largely depends on the quality of the procedures. Competition between various initiatives can trigger an iterative learning process. In this context, it is important to avoid an unnecessary burden on the institutions resulting from multiple

and uncoordinated surveys. Therefore it is desirable that all data obtained for the purpose of comparative performance assessment be made available for alternative assessments and re-analyses in an appropriate and, if necessary, anonymised, form. The data should also be made available for scientific surveys.

The learning process associated with the introduction of comparative assessment procedures need not start from scratch (cf. Section A). For this reason, and on the basis of relevant experience at the national and international levels, the Science Council has issued a number of recommendations on comparative assessment procedures in the system of higher education and research, which should be interpreted as standards for good ranking practice (B.II). Decision-makers in politics and academia alike are called upon to actively support adherence to these standards and, if necessary, voice criticism of any ranking initiatives that fall short of these standards, in order to convince the general public of the quality of rankings.

B.II. Recommendations for comparative assessment procedures in the system of higher education and research

Stating the objectives

In the light of the heterogeneous nature of possible assessment procedures, rankings and related processes[44] require a number of explicit definitions to avoid misinterpretation of results. Clear, user-friendly information must include a statement of objectives and the targeted group of recipients. Whoever uses a ranking – or in most cases, the underlying data – for purposes other than those intended should be cautioned that this may lead to certain problems.

[44] For reasons of brevity, we have sometimes used the term “rankings” instead of “rankings and related processes” or “comparative performance assessments” throughout Section B.II. Thus, all general recommendations refer to all such procedures, unless explicit reference is made to either “rankings in the narrow sense” or “ratings”.

Rankings can have various objectives:

- Supporting scientific institutions in their strategic orientation, profile building and quality assurance;
- Supporting researchers in their self-assessment and strategic planning;
- Promoting competition between institutions;
- Providing assistance for present and future students, doctoral students, young or visiting researchers or for candidates for a professorship in selecting the right institution for the next step in their careers;
- Improving the market efficiency for academic and further training services.

The user groups of rankings in the system of higher education and research vary according to the objective(s) pursued:

- Decision-makers in the institutions themselves;
- Decision-makers at the ministries and within the funding bodies including the foundations;
- Members of universities and non-university research institutions;
- Present and future students, doctoral students, young or visiting researchers and national and international job applicants;
- Companies and public authorities as users of research services.

The objectives and intended users determine the selection of the dimensions of performance, the performance criteria and the objects of the rankings. Furthermore, many aspects of a ranking initiative, including publication channels and intervals between publications, should be organised in such a way as to provide optimum support for user groups without placing an excessive burden on the reviewed institutions.

Selecting decision-relevant object definitions

For each ranking, the set of objects must be defined by naming the scientific discipline, the institutions reviewed and the level of aggregation at which the assessment takes place. Again, the selection is guided by objectives and user groups.

With regard to the categories of institutions reviewed, the Science Council believes that it would be inappropriate to make a direct comparison between universities, technical colleges and non-university research institutions, because these institutions

fulfil very different functions. However, where they provide comparable services, these should be rated on the basis of uniform criteria.

The choice of the level of aggregation is guided by overall objectives and user groups and must also take into account the availability of data and the expense incurred by recording such data.

Rankings that review universities as a whole without any differentiation by subject or function are of questionable value, because of the differences in institutional profiles. Unless they come with clear instructions for interpretation, which in turn presupposes a clearly defined group of intended users, such “league tables” of universities as a whole can easily have undesired effects. By contrast, multidimensional rankings differentiated by subject and function can help institutions create a clearer profile for themselves. The prerequisite, of course, being that such efforts are not countered by ministerial micromanagement.

The Science Council does not recommend public rankings of individuals for purposes other than internal institutional matters. Such rankings would run the risk of being interpreted by many users as a complete and accurate representation of the rated individuals’ academic performance even where such use is expressly advised against. This could lead to the demotivation of researchers that are not among the top ranks and to a violation of their personal rights. Such damage cannot be compensated by the potential incentive provided by public, person-based rankings. Therefore, comparative information on the academic performance of individuals should in principle only be used for internal purposes, such as performance-based allocation of funds. Of course, there is always the possibility of publicly rewarding excellent performance, thus creating incentives and sending clear signals, e.g. to (prospective) doctoral students.

In the case of assessments of institutions differentiated by subject and function at a medium level of aggregation, the object definition must be subdivided into two aspects: Firstly, a classification of scientific subjects, research areas or disciplines must be established. Where possible, this should be done on the basis of one of the classification systems established for statistical or administrative purposes – the DFG’s structure of research areas and subjects or the taxonomy used by the Federal Statis-

tical Office. Secondly, a decision needs to be made whether the empirically existing organisational subdivision of universities and non-university research institutions into institutes, departments, faculties, study courses or other organisational units is to be reflected or whether each researcher working for a reviewed institution is to be individually assigned to a subject/research area.

The assessment of empirically existing organisational units offers the benefit that these objects have already been statistically recorded and are of direct relevance to decision-making. Unfortunately, however, for historic reasons, each university and research institution has its own organisational structure, which makes it more difficult to compare the data obtained. Furthermore, the empirical organisational units are often heterogeneous in terms of subjects, and thus a comparison based on criteria established by the predominant discipline at a given institute or faculty does not do justice to researchers from other fields that are also working for the institution. This type of distortion can be prevented by linking researchers to subjects on the basis of certain norms; however, this comes at a price: the data that have already been gathered within the university and could be relevant as input data for a ranking can only be used to a limited extent.

In the light of these considerations, no generally valid recommendation can be made. As a rule, empirical organisational units at universities are primarily based on the requirements of teaching and should therefore be used as a reference for rankings of teaching performance, whereas research communication and knowledge transfer follow disciplinary or sector-specific structures and are therefore more likely to be adequately recorded in a standard classification system that applies to all organisations (cf. proposal for a research rating in B.III).

Each ranking study should include an explicit definition of its objects. A uniform object definition should be used for all performance dimensions and criteria throughout the study.

Separate assessment of performance dimensions

Universities perform a variety of scientific and science-based services, the most important of which are research, teaching, promotion of young researchers, knowledge transfer, continuing education and (at university clinics) hospital treatment. Neither

are non-university research institutions confined to a single, clear-cut scientific function. In addition, there are other relevant benefits that are not specific to science, such as the promotion of equal opportunities for men and women or the integration of persons with a migration background.

These different services and benefits, which are partly intertwined, and partly differentiated, to a greater or lesser degree, at individual institutions in the course of the differentiation of the system of higher education and research, should be rated independently of each other. It is not advisable to simply squeeze these institutions into a ranking of better or worse. Where a ranking comprises several types of performance, these should be clearly separated as different dimensions of performance and be assessed in separate rankings. Correlations between different dimensions can, if necessary, be illustrated by a multidimensional representation.

Disclosing performance criteria

Within one performance dimension, the assessment can be based on different criteria. Thus, for example, research performance can be rated in terms of the quality achieved at the top level, the (quality-weighted) volume or the efficiency of performance (quantity per unit of input).

The decision on what criteria are used depends in each case on the overall objective and on the interests of the user groups. While it may be important for some young researchers to work with one or just a few particularly renowned colleagues under optimum conditions, it may be crucial for younger students to have a well-qualified teaching staff as a whole. Likewise, higher education policymakers may focus on a university's national and international reputation, whereas, from a fiscal policy point of view, it may be more important to determine where services are provided most efficiently. Differences regarding (implicit) criteria are, for example, likely to be one reason for the fact that ratings of universities by professors typically diverge from ratings by students. Moreover, quality, effectiveness and efficiency do not necessarily always coincide. Adding up or averaging assessments on the basis of what are mutually contradictory criteria would produce meaningless results. Where multidimensional representation allows different criteria to be represented independently of each other, it is possible to serve a larger circle of user groups.

Occasionally, indirect criteria are used, such as resources. Such information on infrastructural conditions leads to expectations of good performance without the quality of the services being substantiated. Nevertheless, this type of information is useful for certain user groups, e.g. young researchers or applicants for a professorship.

Assessment criteria also differ between the various scientific research areas, so that a comparison across subject boundaries or a comparison of units with a heterogeneous structure will lead to bogus results. This applies not only to research, but also to the area of teaching, where expectations regarding graduates also strongly vary from discipline to discipline. The Science Council therefore considers rankings differentiated by subject to be imperative.

Assessing the quality of academic performance requires the involvement of peers from the same discipline to specify and operationalise the general performance criteria of a ranking, since the definition of quality and performance varies greatly between subjects.

Selecting indicators on the basis of performance criteria

Any assessment of academic performance must be based on appropriate quantitative and qualitative indicators. The question of which indicators to use depends on the definition of performance criteria and can thus often only be decided by experts from the discipline concerned. The Science Council does not consider it good ranking practice to merely compile a list of those indicators that can be computed on the basis of easily accessible data. Indicators should not be used if they cannot be clearly assigned to one specific performance criterion.

Wherever possible, the indicators chosen should be compatible with a system of incentives and not susceptible to manipulation. Public ratings generally have a certain influence on people’s behaviour, even where they have no direct bearing on the distribution of resources. This should be taken into account when defining the performance criteria, and even more so when indicator systems are designed. Where it is a known fact that certain indicators tend to induce a certain one-sided optimisation response, they should be used with caution or be modified or combined in such a way as to minimise dysfunctional incentives.

Generally, any ranking initiative should be accompanied by a process in which indicators are reviewed in the light of past experience and, if necessary, modified before being used again.

Aiming at high quality and optimum reusability of data

The data used in rankings must be of optimum quality, in particular because of expected steering effects. Generally, both the data and a description of collection methods have to be published in a manner that meets general scientific standards. In order to ensure transparency and allow control and reusability of the data, the microdata, too, should be made available for re-analysis, taking into account data privacy standards. This can help to limit the overall burden placed by surveys on universities and non-university research institutions.

However, it is important to bear in mind that rankings are not only read by experts. Therefore, good ranking practice requires presenting the surveying methods in a way that is understandable to non-experts, drawing readers' attention to factors that are crucial for judging the reliability of the data and the assessment of the indicators.

Adequate presentation of results

Rankings in the narrow sense lead readers to believe that there are significant differences between the various ranks, not only at the top, but also in the middle and lower ranks. However, such a level of differentiation is usually not required to inform the decisions of user groups. Therefore, it is in most cases preferable to establish ranking groups, each of which is sorted in alphabetical order. If the focus is on the quality of academic output, to be determined by means of a peer-review-based procedure, a rating procedure with predefined quality levels is usually more advisable than a ranking in the narrow sense. Where a ranking involving an ordinal ranking list is considered necessary to achieve certain objectives, it should be examined in each case whether differences with regard to individual indicators that determine the position in the ranking are statistically significant.

Ensuring independent implementation and assessment

The organisers of a ranking should primarily be responsible towards the users and be completely independent of the rated institutions. However, they must enjoy the confidence of both sides.

For nationwide rankings in Germany, the Science Council recommends a significant involvement of foreign academics in order to ensure adequate, internationally valid assessment criteria.

Safeguarding competition and autonomy

Rankings allow documentation of performance differentiation in the system of higher education and research and thus promote competition and profile building. They can be used by universities and non-university research institutions as a source of information to guide development planning and thus as an instrument of self-management.

A distribution of government resources on the basis of a formula that largely depends on ranking results (cf. RAE in the UK) would, it is true, offer the benefit of transparency; nevertheless it would produce a retrospective rather than future-oriented steering effect, since rankings only take into account performance in the past. Rankings are an important source of information for performance-based steering, but should be used in such a way that the autonomy of scientific institutions is reinforced and that there is enough scope for shaping individual profiles. The Science Council therefore advises against a steering mechanism for the system of science and higher education that is completely or predominantly based on ranking results and instead advises decision-makers to use it in combination with other procedures of quality assurance and strategic planning.

B.III. Recommendations for a research rating system

The profile building at universities and their performance differentiation, which is expected to lead to more competition and an orientation towards international standards, nowadays primarily take place in the area of research. Non-university research institutions play an important role in the German research landscape. Universities and non-university research institutions are mutually complementary in their

missions and should bear that in mind in their strategic decision-making. Reliable comparative quality information is needed to support them in this process and to make competition in research more effective and efficient. Therefore the Science Council recommends a nationwide comparative assessment of research performance which rates institutions on the basis of international criteria. A research area-specific, multidimensional assessment based on various criteria is needed to support universities in the profile building process.

In the light of international experience (cf. A.II and the resumé in A.IV), the Science Council rules out research assessment systems that are solely based on quantitative indicators, as well as mere reputation-based ratings. A comparison of the quality of research performance requires a research area-specific assessment in the form of a peer review carried out on the basis of harmonised data and quantitative indicators ("informed peer review") on a predefined assessment scale. In that sense, the procedure recommended by the Science Council is a research rating.

The overall concept recommended for a research rating includes the following components: a steering group (cf. III.7) establishes assessment panels (III.5) that rate the performance of the reviewed institutions in their respective research areas along the dimensions of research, promotion of young researchers and knowledge transfer according to generally defined criteria later detailed by each panel (III.2). For this purpose, they are presented with the institutions' research area-specific profiles (III.3) and quantitative indicators (III.4). The results are shown as ranking groups of universities and non-university institutions for the various research areas (III.6). In addition, quality profiles are to be established for each institution. For the medium term, there are plans to examine the possibility of establishing an international benchmarking system in cooperation with other countries in which comparable procedures are in place (III.9).

III.1. Objectives, intended users, objects

The objective of the research rating recommended by the Science Council is to provide support for universities and non-university research institutions both in their missions and – in connection with other quality assurance and strategic planning procedures – in quality assurance measures in research, and to promote competition for

quality. For this purpose, comparative information on their research performance is to be provided and rated according to international standards. Important aspects that are closely related to research performance include promotion of young researchers and knowledge transfer. They are of decisive importance for the contribution of each institution to the successful practice of science in Germany and therefore also need to be rated in order to boost competition in these dimensions, too.

The intended users of the research rating are decision-makers at the universities, both on the university governing boards and at the level of faculties and institutes, as well as at non-university research institutions and in the competent ministries. Cooperation between these levels is expected to boost competition in the research arena with a view to achieving a durable improvement in quality and allowing cutting-edge outputs. The transparency achieved through the research rating is an important prerequisite for this in that it helps to avoid local standards.

However, the research rating will be of interest not only for decision-makers within the various organisations, but also for undergraduate students, doctoral students and young researchers. The same applies to visiting scientists and applicants for a professorship, although for that group, other, more individual channels of information are likely to play a more important role.

The research rating targets research activities of universities and non-university institutions; the latter should be listed separately as a category in their own right. Service institutions in the research sector are not covered by the rating.

A research rating must be differentiated by subject, because the various subjects and research areas are governed by different quality criteria. The Science Council recommends that the steering group (cf. III.7) which is entrusted with defining the final taxonomy of research areas in consultation with subject representatives should elaborate the new system on the basis of the DFG’s current taxonomy of research areas, review boards and subjects. This will allow the collected data to be used more than once. In individual cases, it may be advisable to use groups of subjects for the purposes of the rating or introduce further subdivisions along the subject boundaries defined by the DFG classification. The number of research areas in the final taxonomy should not exceed 50.

The ratings in their entirety add up to a certain quality profile of research outputs for each institution's various research areas. The Science Council acknowledges the possibility of using such quality profiles to produce a ranking of universities and non-university research institutions in Germany. However, this presupposes a weighting of research areas and performance dimensions that depends on users' individual interests; therefore the Science Council does not recommend introducing a general formula for such a weighting. If the aim is to produce an institution ranking for various subjects on the basis of a research rating, interpretation of the ranking should involve use of information on the mission, research strategy and organisational structure of the entire institution, as well as more highly aggregated data, for example on institutions' budgets.

III.2. General performance dimensions and assessment criteria

The research rating is designed to comprise assessments of universities and non-university research institutions along the three performance dimensions of research, promotion of young researchers and knowledge transfer and specifically on the basis of a number of criteria which will be described in general terms in the following. The specification and operationalisation of these assessment criteria and the selection of criteria as such are subject-specific and contingent upon the availability of reliable data. This task is performed by the assessment panels to be established for the various research areas (III.5). The process is based on research profiles submitted by the organisations (III.3) and on quantitative indicators (III.4).

The research dimension

The central criterion in the research dimension is quality. The other relevant criteria, i.e. impact and efficiency, are linked to the aspect of quality; however, their separate assessment allows assessment panels a differentiated judgement in each individual case.

First criterion: quality – This criterion comprises the topicality and relevance of the topics for the research area in question as well as the novelty and originality of the research results and the suitability and reliability of methods. Data on the response to the results from peers (e.g. standardised relative citation indicators) can be used as indicators.

Second criterion: impact – This criterion assesses the reviewed institution's contribution to the development of the research area in question, if measured by international standards. The quality-weighted number of research outputs should be measured by suitable indicators.

Third criterion: efficiency – Here the quality-weighted quantity of research outputs as required by the criterion of impact must be rated in relation to the research-related input. This input can be estimated using aspects such as deployment of human and financial resources.

The dimension of promoting young researchers

Fourth criterion: Processes for promoting young researchers – This criterion is about assessing the measures taken by the institution to promote young researchers. Indicators include the establishment of structured doctoral programmes, average duration of doctoral studies (median) and the number of young researchers in independent scientific positions.

Fifth criterion: Successful promotion of young researchers – This criterion revolves around a quality-weighted assessment of the contribution provided by the institution during the period of assessment to promote young researchers in a given research area. The academic success of its graduates can, for example, be measured by looking at their publications (and critical response from peers). A broader indicator of success is the proportion of postgraduates who obtain an adequate position. The spectrum of occupational opportunities considered adequate differs from one research area to another. Apart from careers in the academic world, engineering and managerial positions in business, in the cultural and educational arenas and in politics and administration are considered satisfactory. However, such an assessment comes with a long delay.

At present, only few universities in Germany are in a position to provide reliable information about the success of their graduates, so that survey results in the first round should be treated with caution. Nevertheless, the Science Council considers it indispensable to substantially improve the level of information on the career paths of young researchers, in order to be able to rate the quality of the institutions' promotion measures for young researchers and make the training and labour market for young

researchers more transparent. The Council therefore advocates gathering data on this aspect[45] and is hopeful that this will provide a strong incentive for universities to document the success of their promotion measures for young researchers by improving their alumni support programmes and by carrying out studies on graduates' careers.

In the same way, non-university research institutions should document the professional careers of their doctoral students and young researchers.

The knowledge transfer dimension[46]

Sixth criterion: relevance – Here the question is whether the research results are of relevance for scientific progress in other disciplines beyond the confines of the research area concerned, and of practical relevance.

Seventh criterion: application in business – The application of relevant research results in new products or services is a crucial criterion for measuring success, particularly in practice-oriented fields.

Eighth criterion: further and continuing education – Further and continuing education courses can be an important vehicle for universities to spread research-generated knowledge in society. Here evaluation should include not only qualitative and quantitative aspects, but also the model character of measures.

Ninth criterion: research-based consulting, contributions to public understanding of science and humanities – In many research areas, research-based consulting services are a central source of knowledge transfer for private businesses and public authorities. Imparting scientific methods and findings to the public (*Wissen-*

[45] The assessment panels should also make use of existing studies on graduates' careers that were not carried out by the universities themselves, but by higher education researchers. However, it is usually not possible to compare universities on the basis of these studies. Thus, in the Career after Higher Education: a European Research Study (CHEERS), which surveyed 40,000 graduates, the institution by which the degree was awarded was not one of the variables examined (cf. http://www.uni-kassel.de/wz1/TSEREGS/metho_e.htm and the literature listed therein).

[46] In this performance dimension, the interpretation of the underlying criteria and the availability of data on which the assessment can be based differ from subject to subject. It should be up to the assessment panels to choose the appropriate ones among criteria 6 to 9 and complement them where necessary.

schaft im Dialog/PUSH, exhibitions, etc.) is an important type of knowledge transfer provided by universities and non-university research institutions.

III.3. Research profiles of institutions

The assessment is based on institutions' research profiles, including both qualitative and quantitative data, as well as on other quantitative indicators derived from existing data (B.III.4). Each institution submits a profile for each of its research areas. Thus, depending on its range of subjects, each institution can submit up to approx. 50 research profiles.[47]

Each research profile should be preceded by a brief introduction outlining the role of the research area within the institution's overall strategy ("mission") as well as the specific strategy pursued in that research area. This should be followed by a standardised documentation of research activities in the research area concerned (cf. Annex), which includes a basis for the assessment of each assessment criterion. This basis can be complemented by separately recorded quantitative indicators (in italics, cf. III.4).

If so requested by the assessment panels, other types of information going beyond the standardised research profile can be recorded, provided that this does not involve an unreasonable expense for the reviewed institutions. Moreover, institutions should be given the option of transferring any further information they consider indispensable for the assessment (stating research-based activities that are not covered by the standard research profile).

Assessment criteria (III.2)	**Basis of assessment in the research profile (Annex)**
Research dimension	
1. Quality	Research outputs (2.) Third party funding (4.a) Scientific cooperation projects (8.) *If applicable, relative citation indicators, proportion of highly-cited publications*

[47] In specific cases, an institution may submit more than one research profile for a certain research area. However, this needs to be agreed with the assessment panel in advance.

2. Impact	Quality-weighted publication figures (3.a/b) If *applicable*, *presentations at major international conferences* *If applicable, absolute citation counts, impact-weighted publication counts*
3. Efficiency	Numerator: quality-weighted publication figures (3.a/b) Denominator: Number of researchers (1.b), Resource input including third party funding (4.)
Dimension of promoting young researchers	
4. Processes for promoting young researchers	Structured doctoral programmes, median of duration of doctoral studies (5.) Externally funded fellowships (4.) Number of independent junior research groups
5. Successful promotion of young researchers	Subsequent career of doctoral students, postdoctoral students (5., 6.) Publications of young researchers (3.a)
Knowledge transfer dimension	
6. Relevance	Research outputs (2.) Cooperation projects (8.)
7. Application in business	Funds from industry (3.) Intellectual property rights, licences (7.) Start-ups (7.) Cooperation projects (8.)
8. Training and further training	Description of training and further training measures (9.)
9. Research-based consulting, scientific communication	Description of research-based consulting services and science communication activities

Table: Assignment of components of research profiles to assessment criteria

III.4. Quantitative indicators

Quantitative indicators, and especially bibliometric indicators, for measuring the volume and quality of research outputs are now recognised in many research areas. The assessment panels should make use of such indicators, provided that the relevant prerequisites in their research areas are met. In the research dimension, the following indicators are most important:

- Absolute publication and citation figures used as a measure of the impact of an institution in the research area in question;[48]
- Publication figures weighted with citations (or alternatively/in addition, a quality factor for the journal in question), used as an indicator of impact;[49]
- A relative citation measure standardised by subject, used as a quality indicator;[50]
- A quotient corresponding to the proportion of particularly highly cited papers (top percentile) in the research area concerned, used as a quality indicator.[51]

Since the purpose of the research rating is to compare the current research potential of the reviewed institutions using data that is as recent as possible, bibliometric data should be recorded using the current potential method (on the basis of the names of the researchers employed by the institution as at the reference date as listed in the research profile).

The conduct of bibliometric analyses in the research areas for which there are suitable databases and for which a time frame of five years is sufficient, must be duly taken into account in the schedule and the funding plan for the research rating. The reviewed institutions should cooperate with the institute carrying out the analysis in updating authors' addresses in the databases used.

III.5. Research-area-specific assessment by assessment panels

The assessment is carried out by research-area-specific assessment panels, each of which is composed of up to six researchers, including two researchers from abroad, and another two experts from outside the realm of research institutions funded by public authorities (depending on the research area: industry, culture and education,

[48] For the purpose of the rating, empirical citation data should only be used in research areas in which a time frame of no more than five years (this applies to the oldest publications, whereas more recent publications are subject to shorter periods) is sufficient to make a reliable statement about the impact of the publications.

[49] Due to the skewed distribution of citations, a journal's impact factor, which is frequently used for weighting, says little about the quality of individual publications contained in it. However, at the institutional level of the research rating, with a certain size of samples, the use of a quality factor modelled on the impact factor can be useful.

[50] Citations per publication (= relative citation factor), divided by (=standardised by) the *average* number of citations per publication *in the research area*. In a standardised measure, a value of 1 shows an average, and a value above 1, an above-average level of reception for the publications of the reviewed institution.

[51] Tijssen, Visser & van Leeuwen (2002); for a comparison with traditional bibliometric indicators, see van Leeuwen et al. (2003). For this type of indicator, merely methodological papers should be given a lower weighting.

politics, administration), as well as an observer not representing the research area. The size and composition of the various groups depend on the breadth of the research area concerned and are determined by the steering group.

Each observer attends the consultative meetings of four to five assessment panels. His/her task is to ensure uniform assessment standards.

At its constitutive meeting, each assessment panel agrees on how to interpret the assessment criteria, what indicators to use for evaluating them and how to weight these indicators. A statement outlining these decisions is published and sent to the institutions being reviewed.

The assessments are based on the research profiles to be submitted for each research area by universities and non-university institutions (III.3 and Annex). In addition, the assessment panels should use quantitative indicators that are already available or can be obtained at a reasonable expense in order to substantiate and verify their judgement, provided that this is in line with established practice in their subjects (III.4). In difficult cases, the assessment panels should be free to hold meetings with representatives of individual institutions or visit their premises.

The assessment panels are requested to rate the research profiles on a seven-point scale according to each of the criteria listed above. For the main criterion, i.e. “research quality”, the key values on the scale are given a verbal definition:

7	More than half of research activities are of top international standard; all other activities are of top national standard and are internationally competitive.
5	More than half of research activities are of top national standard and are internationally competitive. Some activities may be of top international standard.
3	The research activities for the most part conform to national quality standards. Some activities may be of top national standard and be internationally competitive.
1	The research activities do not, or only in a few cases, conform to national quality standards.

The numbers between these predefined values are designed to allow a fine-tuning of the assessment. The assessment panels are requested to make use of the full range of the scale. If international standards of excellence are adhered to, the best mark will not be awarded in each research area.

III.6. Presentation of results

The rating results in multidimensional assessments of the research activities of universities and non-university institutions for each of the reviewed research areas. The adequate form of publication is a web-based, dynamic publication offering two different retrieval functions.

On the one hand, users should be enabled to sort the activities of the various universities in a given research area according to any of the nine predefined assessment criteria. The results based on the chosen criterion should be shown in seven ranking groups (as defined by the scale), in which universities are sorted in alphabetical order. Likewise, such ranking groups can be established for non-university research institutions in any given research area, according to appropriate criteria commensurate with their respective missions.

On the other hand, it should be possible to choose a certain university or non-university institution and retrieve all assessments that refer to its research activities as a whole. This leads to a strategically useful representation of the institution’s overall research portfolio.

The basic data obtained from institutions on personnel structure, third party funding and doctorates should also be published.

III.7. Implementation, organisation and funding

The research rating should be carried out under the aegis of an organisation that is linked to the academic community and has the confidence both of universities and non-university institutions and of the federal and state governments. This body should also possess organisational and methodological competence in research assessment and be independent of the reviewed institutions.

Coordinating responsibility for the research rating must lie with a steering group composed of renowned researchers from all fields of science and humanities. Furthermore, the major research organisations should be adequately represented. The role of the federal and state governments within the steering group will be laid down definitively when the pilot study (B.III.8) has been concluded.

The central tasks of the steering group are:

- Define the taxonomy of research areas in consultation with the research societies
- Appoint the chairpersons of the assessment panels
- Appoint the members of the assessment panels
- Adopt guidelines for the creation of research profiles
- Adopt guidelines for the work of the assessment panels
- Ensure uniform, consistent assessment criteria
- Assume overall responsibility for the research rating.

The steering group and the assessment panels must be supported by a secretariat, which, in administrative terms, should be part of the implementing organisation, but, in its work, should be solely responsible to the steering group.

The creation of research profiles at universities and non-university institutions should take place electronically. The university governing boards should be able to reexamine the research profiles created for the rating before they are submitted to the assessment panels. If possible and achievable at a reasonable expense, quantitative data should be validated using external sources. The results of the survey should be made available for other evaluations and be presented in a user-friendly way. When the first phase of the buildup of the planned Institute for Research Information and Quality Assurance (IFQ)[52] has been concluded, it should be examined whether the institute can be requested to collect data and evaluate existing data sources (third party funding statistics of donors, bibliometrics) for the rating and whether it can provide an input to the development of quantitative indicators.

[52] The *Deutsche Forschungsgemeinschaft* has created this institute as an auxiliary institution, in order to process information on its funding activities and other data on the system of higher education and research in general in such a way as to enable DFG, in a first phase, to improve its funding activities and, in a second phase, to provide support for other players in the system of higher education and research.

Assessment in a research area will take approximately 15 to 18 months from the appointment of members of the assessment panel via the operationalisation of criteria and the creation of research profiles to the publication of results. For organisational reasons and because of the burden placed on universities and non-university research institutions, it is hardly feasible to assess all research areas (approx. 50) simultaneously. Since the purpose of the rating, unlike in the case of the RAE in the UK, is not to guide the decision-making process with regard to basic funding in the higher education sector, such a simultaneous evaluation is not an absolute necessity. Instead, a rolling schedule should be adopted under which a certain number of research areas are reviewed each year.

The process needs to be repeated at regular intervals if the lessons from the research rating are to be learned. International experience suggests that intervals of five to six years are recommendable. If 12 to 15 research areas are assessed each year, all disciplines will be covered within four years, which leaves roughly a year for summarising the results, evaluating the process and making adjustments, if necessary.

The universities and non-university institutions are likely to need a substantial amount of clarification with regard to both the objectives and implementation of the research rating. Therefore a comprehensive communication strategy is required which should include information events, a help function on the website of the rating and telephone consulting. In addition, the Science Council recommends holding regional conferences during the implementation of the research rating in order to make universities and non-university institutions acquainted with the objectives and methods used in the process. This is indispensable to ensure the quality of the resulting data and the acceptance of the research rating altogether.

The direct costs incurred through the implementation of the research rating include fees and travel expenses for the reviewers, the cost of bibliometric and other database analyses, the cost of creating the research profiles and the staff costs for the project leaders, the support for the steering group and the assessment panels as well as for PR work. The Science Council estimates that these costs will altogether amount to at least € 2.6 million a year.

Furthermore, a significant amount of indirect costs is expected to be incurred by the reviewed institutions in creating the standardised research profiles and collecting the necessary data. These costs may exceed direct costs several times over.

Given the great importance of a reliable, methodologically ambitious research rating for scientific competition in Germany, the Science Council considers this expense to be justifiable.

III.8. Pilot study

The Science Council recommends testing the suitability of the method proposed here in a pilot study. Apart from numerous operative aspects, a number of more general questions cannot be clarified without such a study:

- It must be examined what level of detail is required for the taxonomy and the list of criteria in order to achieve the objectives of the research rating. The aim should be to make the procedure as simple as possible.
- The criteria must be specified and operationalised according to subject, without, however, abandoning the uniform nature of the procedure altogether.
- The possibility of further aggregating the information by weighting the criteria as prescribed should be examined.
- It should be examined whether part of the data could be updated at shorter intervals in order to provide users with more up-to-date information.
- Appropriate rules must be established for the assessment of interdisciplinary research units and institutes in order to obtain adequate results despite the primarily subject-based structure of the research rating.
- Where universities and non-university research institutions have evaluation procedures in place, it needs to be examined whether any current data collected for such procedures can be used for the rating, in order to keep survey costs to a minimum.
- The cost/benefit ratio of the procedure must be determined.

For the purposes of the pilot study, two research areas should be chosen which differ significantly from each other in methodological terms and which allow as many of the anticipated problems as possible to be studied. At the same time, these should be research areas whose delimitation is uncontroversial, so that the results can be

placed in the taxonomy defined for the actual research rating. For example, informatics and sociology would be two suitable subjects.

In order to allow the pilot study to begin as soon as possible after adoption of these recommendations, a working group of the Science Council should be requested to operationalise the procedure and appoint the assessment panels. This working group could at the same time be the core of the future steering group, thus ensuring the continuity of the procedure. The major scientific organisations should already be involved in the process at the stage of the pilot study. Representatives of the federal and state governments should have a guest status in the working group during the pilot phase.

Even during the pilot study, it is imperative to pursue an active information and communication policy to improve understanding of the procedure among decision-makers at universities and non-university institutions and ensure optimum coordination during the collection of data.

The steering group should evaluate the experience made during the pilot study and report its findings to the Science Council, which in turn will consult the scientific organisations. The Science Council reserves the right to state its opinion on the implementation of a future research rating system after receiving the results of the pilot study.

III.9. International benchmarking

The proposed research rating system allows a comparative assessment of the research performance of universities and non-university institutions in Germany. The assessment panels are required to apply international standards. This is to be achieved through various aspects of the procedure: international composition of assessment panels, definition of the assessment scale and resulting instructions to reviewers, and use of internationally standardised indicators.

An international benchmarking of universities and non-university research institutions can only be carried out in cooperation with other countries, because the necessary data is not freely accessible. As well as allowing mutual control of assessment stan-

dards, such cooperation can also help all parties involved to learn from each other in methodological terms.

The procedure proposed by the Science Council shows similarities with the assessment procedures used in the Netherlands and in the United Kingdom of Great Britain and Northern Ireland. Therefore the Science Council recommends examining, together with representatives of the Netherlands and the UK, ways of implementing a benchmarking system for assessment procedures and mutual controls of assessment criteria in the three countries once the pilot study for the research rating has been completed. This would enable the three countries to obtain a more reliable picture of the standing of their research institutions.

Annex

References

Bayer, Ch. R. (1999): “Hochschulranking,. Übersicht und Methodenkritik”; published in *Beiträge zur Hochschulforschung*, special issue.

Berghoff et al. (2003a): *Das Hochschulranking. Vorgehensweise und Indikatoren*; Centrum für Hochschulentwicklung, Arbeitspapier 46, Gütersloh.

Berghoff et al. (2003b): *Das CHE-Forschungsranking deutscher Universitäten 2003*; Centrum für Hochschulentwicklung, Gütersloh.

Berghoff S. & S. Hornbostel (2003): “Das CHE hinter den Sieben Bergen”; published in *Perspektiven der Wirtschaftspolitik* 4, p. 191-195.

Bornmann, L. & H.-D. Daniel (2003): “Begutachtung durch Fachkollegen in der Wissenschaft – Stand der Forschung zur Reliabilität, Fairness und Validität des Peer-Review-Verfahrens”; published in S. Schwarz & U. Teichler (eds.), *Universität auf dem Prüfstand – Konzepte und Befunde der Hochschulforschung.* Campus Verlag, Frankfurt/New York, p. 207-225.

Büttner, Th., M. Kraus & J. Rincke (2002): *Hochschulranglisten als Qualitätsindikatoren im Wettbewerb der Hochschulen*; ZEW Discussion Paper 02-78, Mannheim.

Carnegie Foundation for the Advancement of Teaching (2001): *The Carnegie classification of institutions of higher education. 2000 edition*. Menlo Park/Ca.

Centrum für Hochschulentwicklung (2002): “CHE-Forschungsranking. Forschungsstarke Fakultäten an deutschen Universitäten.“ published in *DUZ – das unabhängige Hochschulmagazin*, supplement of 8 November 2002.

CEWS – Kompetenzzentrum Frauen in Wissenschaft und Forschung (2003): *Hochschulranking nach Gleichstellungsaspekten*. cews.publik.no.5, Bonn.

Daniel, H.-D. (1988): “Forschungsleistungen wissenschaftlicher Hochschulen im Vergleich”; published in Daniel & Fisch (eds.), p. 93-104.

Daniel, H.-D. (2001): “Was bewirken Hochschulrankings? Wer orientiert sich an ihnen?”; published in D. Müller-Böling et al. (eds. 2001), p. 121-124.

Daniel, H.-D. & R. Fisch (eds. 1988): *Evaluation von Forschung*. Universitätsverlag Konstanz. Konstanz.

Deutsche Forschungsgemeinschaft (1997): *Bewilligungen nach Hochschulen. Bewilligungsvolumen 1991 bis 1995. Anzahl kooperativer Projekte im Jahr 1996*. Bonn.

Deutsche Forschungsgemeinschaft (2003): *Förder-Ranking 2003. Institutionen – Regionen – Netzwerke*. Bonn.

Engel, U. (ed. 2001): *Hochschul-Ranking. Zur Qualitätsbewertung von Studium und Lehre*. Campus Verlag, Frankfurt/M.

Fabel O., E. Lehmann & S. Warning (2002): “Der relative Vorteil deutscher wirtschaftswissen-schaftlicher Fachbereiche im Wettbewerb um studentischen Zuspruch: Qualität des Studiengangs oder des Studienstandortes?”; published in *Zeitsschrift für betriebswirtschaftliche Forschung* 54, p. 509-526.

Goldberger, M. L., B. A. Maher & P. E. Flattau (eds. 1995): *Research-Doctorate Programs in the United States. Continuity and Change*. National Academy Press, Washington, D.C.

HEFCE – Higher Education Funding Council for England (2003): *Joint Consultation on the review of research assessment*. Bristol.

Hornbostel, S. (1997): *Wissenschaftsindikatoren. Bewertungen in der Wissenschaft*. Westdeutscher Verlag, Opladen.

Hornbostel, S. (2001): “Der Studienführer des CHE – ein multidimensionales Ranking”, published in Engel (ed.), p. 83-120.

Jones, L. V., G. Lindzey & P. E. Coggeshall (eds. 1982): *An Assessment of Research-Doctorate Programs in the United States*. National Academy Press, Washington, D.C.

Leeuwen, Th. N. van, M. S. Visser, H. F. Moed, T. J. Nederhof, A. F. J. van Raan (2003): “The Holy Grail of Science Policy: Exploring and combining bibliometric tools in search of scientific excellence”. *Scientometrics* 57, p. 257-280.

Leszczensky, M. & F. Dölle (2003): *Ausstattungs-, Kosten- und Leistungsvergleich von Universitäten. Ergebnisse für Norddeutschland und Berlin für das Jahr 2000.* HIS Hochschul-Informations-GmbH, Hannover.

Max-Planck-Gesellschaft (2002): *Evaluation. Die Verfahren der Max-Planck-Gesellschaft*. MPG, München.

Müller-Böling, D. et al. (eds. 2001): *Hochschulranking. Aussagefähigkeit, Methoden, Probleme*. Verlag Bertelsmann Stiftung, Güterloh.

NWO, VSNU & KNAW (2002): *Standard Evaluation Protocol 2003-2009. For public research organisations*. http://www.vsnu.nl/show?id=42615&langid=246

Ostriker, J. P. & Ch. Kuh (eds. 2003): *Assessing Research-Doctorate Programs. A Methodology Study*. The National Academy Press, Washington, D.C.

Quality Assurance Agency for Higher Education (2000): *Handbook for academic review*. QAA, Gloucester.

Raan, A. F. J. van (1996): “Advanced Bibliometric Methods as Quantitative Core of Peer Review Based Evaluation and Foresight Exercises”. *Scientometrics* 36, p. 397-420.

Research Assessment Exercise (2001): *A Guide to the 2001 Research Assessment Exercise*. www.hero.ac.uk/rae/pubs

Research Assessment Exercise (2004): *RAE 2008. Initial decisions by the UK funding bodies*. Bristol, February 2004.

Rosigkeit, A. (1997): “Hochschul-Ranking. Hintergründe und kritische Anmerkungen zu einem modernen Bewertungsverfahren”; published in *Beiträge zur Hochschulforschung* 1, p.23-49.

Stern (2003): *Der Studienführer 2003.* Stern Spezial Campus und Karriere 1/2003.

Tijssen, R. J. W., Th. N. van Leeuwen & A. F. J. van Raan (2002): *Mapping the Scientific Performance of German Medical Research*. Schattauer Verlag, Stuttgart.

Tijssen, R. J. W., S. Visser & Th. N. van Leeuwen (2002): “Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference?” *Scientometrics* 54, p. 381-397.

Ursprung, H. W. (2003): “Schneewittchen im Land der Klapperschlangen: Evaluation eines Evaluators”; published in *Perspektiven der Wirtschaftspolitik* 4, p. 177-189.

Wissenschaftliche Kommission Niedersachsen (2002): *Forschungsevaluation an niedersächsischen Hochschulen und Forschungseinrichtungen. Grundzüge des Verfahrens*. Version published on 20 December 2002.

Wissenschaftsrat (1985): *Empfehlungen zum Wettbewerb im deutschen Hochschulsystem*. Köln.

Format of research profiles

The form and content of research profiles should be determined during the pilot study and be defined by the steering group as the research rating progresses, with subject-specific exceptions being admissible if recommended by the assessment panels. The following pattern will illustrate the recommendations of the Science Council:

Research profile of the institution ... in the research area of ...

Period of assessment: ... to ... (last five years)

Reference date: ...

Brief information on the mission of the institution, the role of the research area within that mission and the institution's strategy for the research area.

1. *Scientific staff*

a) List of names of scientific staff for the surveyed period, marking the staff employed as at the reference date, split by staff category and source of funding (basic funding/ third party funding), specifying, if applicable, any part-time employment.

b) Number of staff by category and source of funding in tabular form (in full-time equivalents)

c) List of working units (professorships / institutes / departments) and their staff (optional)

Explanation: The evaluation is carried out independently of this correlation for the entire range of research activities in the research area. However, it is possible to apportion individual working units to several research areas. This has an effect on the rating of efficiency.

2. Selected research outputs: List of selected research outputs produced during the surveyed period. These can be books, articles in books or scientific journals, conference presentations, exhibitions or exhibition catalogues, multimedia productions or other outputs, provided that they are recognised as research outputs in their respective disciplines and have been published before the closing date.

A minimum of three research outputs must be named for each research area. If more than three researchers are working in that research area, one more research output must be presented for each commenced three researchers.

The institutions agree to provide the assessment panel with a copy of each listed research output on request.

3. Publications:

a) Complete list of publications from the surveyed period, split by type of medium (monographs, features in edited volumes, papers in refereed and non-refereed journals, self-published publications, electronic publications, presentations at major international conferences); marking of publications that are the result of dissertations.
b) Tabular list of publication counts from the surveyed period, split by type of medium.

4. *Resource input*

a) Third party funds spent: Tabular list of third party funds spent annually during the surveyed period, split by donors (DFG, Federal Government, state governments, EU,

foundations, industry, other); list of externally funded fellowships.
b) Basic funds spent (if data is available for individual subjects)

Supervised doctoral theses: Number of doctoral students as at the reference date; number of doctorates awarded during the surveyed period; median of duration of doctoral studies; whereabouts of PhD holders one year after completion of doctoral studies (academia, industry, media, administration, unemployment, unknown; in each case: home or abroad); scope, funding and profile of structured doctoral programmes.

Postdocs: Number of postdocs by category (scientific staff, heads of independent junior research groups, Emmy Noether fellows – see list under 1.); number of persons leaving the institution during the surveyed period; whereabouts of these persons one year after leaving (professorship, other academic position, industry, media, administration, unemployment, unknown; in each case: home or abroad).

Transfer to industry: Income from contract research (see 3.), intellectual property rights and licences that are a direct result of research activities, start-ups launched by members of the institution or by the institution itself and number of equity participations.

Cooperation projects: Specification of most important cooperation partners from academia and business, including data on the contractual and financial basis of cooperation projects; participation in projects receiving combined funding.

Training and further training: Brief description of training and further training opportunities offered in the research area, including data on curriculum, number and type of participants, fees, resources, rating by users.

Research-based consulting and scientific communication: Data on consulting services provided for industry, politics and administration (number of contracts, revenue); data on activities of scientific communication (exhibitions, open-house events, *Schüler-Universität*, *Wissenschaft im Dialog*, etc.).