WISSENSCHAFTSRAT

HEAD OFFICE

Drs. 8893-08
Cologne, 11.12.2008

Steering Group Report on the Pilot Study Research Rating in Chemistry and Sociology

Inhalt Seite

Preface

In November 2004, The German Council of Science and Humanities [in the following also referred to as the Wissenschaftsrat or the Council] presented recommendations for rankings in the scientific research system. Part of that publication was dedicated to a methodical review of the existing ranking schemes and the development of a procedure for research rating.[1] To test this procedure, the Council of Science and Humanities decided, in July 2005, to conduct a pilot study for two disciplines, chemistry and sociology. The Council commissioned a steering group composed of members of the Council, representatives of scientific organizations and other experts to carry out the study. This steering group, in turn, appointed assessment boards for each of the two disciplines, including national and international experts in the respective subjects.

The present report contains the steering group's characterization and assessment of the implementation of the pilot study on research rating in chemistry and sociology, and recommendations for the future of the research rating system. It is based, essentially, on the experiences of the two assessment boards, documented in their respective final reports[2].

The steering group adopted this report on the pilot study Research Rating in Chemistry and Sociology on April 10, 2008.

---

1 Wissenschaftsrat: Recommendations for rankings in the system of higher education and research. Part 1: Research. For an english translation see www.wissenschaftsrat.de..

2 Pilotstudie Forschungsrating Chemie: Abschlussbericht der Bewertungsgruppe (Wissenschaftsrats, Drs. 8370-08) and Pilotstudie Forschungsrating Soziologie: Abschlussbericht der Bewertungsgruppe (Wissenschaftsrat, Drs. 8422-08.

Summary

In November 2004, the German Council of Science and Humanities published the concept for a research rating procedure and recommended testing this procedure through a pilot study. According to a decision made in July 2005, the pilot study was conducted for the disciplines of chemistry and sociology. The study was taken up in fall 2005 and concluded early 2008.

To implement the pilot study, the Council commissioned a steering group, which in turn appointed expert reviewers to two assessment boards, one for each of the disciplines evaluated in the study. The present report summarizes the steering group's experiences gained from the pilot study and is partly based on the final reports of the two assessment boards.

Compared to conventional ranking systems, the novel research rating stands out by a number of unique characteristics:

- The quality of research is assessed by an "informed peer review" procedure based on quantitative and qualitative comparative data and taking into account context information.
- The relevant learned societies contribute to the definition and operationalization of the assessment criteria.
- The documented differentiation of the quality of research within the respective institutions enhances the informative value of the results.
- The rating by a range of criteria ensures that the results reflect the variety of performance profiles of different research institutions.
- By including non-university research institutions, which play an important role in disciplines such as chemistry, the rating exercise provides a comprehensive picture of Germany's research landscape.

Although the quality of research remains the core criterion, the relevance of the assessment in terms of each of the criteria research quality, impact/effectiveness, efficiency, promotion of young researchers, knowledge transfer to other areas of society and promotion of the public understanding of science also depends on the scope and mission of each individual institution, which, therefore, must be taken into account when interpreting the ratings.

The pilot study has shown that the research rating system can be adapted to disciplines with very different research practices. For instance, in chemistry the assessment of the core criterion, research quality, is partly based on citation indicators, whereas in sociology, due to the heterogeneous practices of publication and the resulting shortcomings of the data, the assessment procedure had to do without such indicators. Instead, in sociology the assessment of the research quality is based, to a considerable extent, on the reading of selected publications. Both methods resulted in assessments that are differentiated as well as reliable, as evidenced by the high level of agreement between reviewers.

The progress and results of the pilot study justify the expectation that research rating can be applied successfully in other disciplines, too. It is the steering group's view that the procedure should be developed further, step by step. This development should also include further improvements of the definition and data basis for some criteria. In particular, it should be examined whether differences in the workload from tasks other than research, e.g. teaching, disparities in the research infrastructures available, and variations in the resources required by different branches of the respective disciplines can be taken into account in more detail at acceptable cost. It would also be desirable to have the criteria in the knowledge transfer dimension specified more precisely for individual disciplines, and to improve the data basis in this respect.

The costs for the chemistry and sociology research rating exercises was considerable, but appropriate for a pilot study. For any further development of the research rating system, the expense must be kept within acceptable limits, and the informative value of the results must be optimized. The steering group recommends:

- further standardizing the definition of the research units;
- reducing the volume of data to be collected;
- increasing the lead time for data collection, to make it easier for research institutions to prepare for the assessments;
- standardizing, as far as possible, the data collection formats with other organizations collecting data, so that data will be suitable for multiple use;
- optimizing the data quality, to reduce the workload of the assessment boards.

Should the procedure for research rating be developed in this direction, the steering group recommends next to assess one discipline from the humanities and one from the technical sciences, since these are the fields most different from natural and social sciences, both in their internal publication and communication channels and in their relations to other areas of society, which are relevant for the transfer dimension of research.

The decision about the permanent establishment of research rating should be preceded by further clarification, in dialog with the users, of the benefits from the published results of the pilot study, including a comparison with any evaluations carried out by the research institutions themselves. Furthermore, the Council should support an examination of the consequences of evaluation procedures like this.

A. Initial situation

A.I. Science policy context

The system of scientific institutions in Germany has changed significantly over recent years. It has become accepted wisdom that the university system can meet the growing demand for tertiary education, and achieve a research performance of international excellence, and support businesses by practice-based research and training provisions, only if the individual universities concentrate on focus areas and, by competing with each other, develop differentiated university profiles.[3] Today, such differentiation is actively promoted by German Federal and Länder authorities, for instance, paradigmatically, in the Excellence Initiative. At the same time, in the context of this competitive differentiation, rigid delimitations between the university and non-university research sectors are increasingly broken up by collaborations, which, over recent years, have become ever more numerous, intensive and binding.

An important factor in these changes is that the relation between the state and the sciences was transformed from a classic bureaucratic system to a more competitive model. Target agreements are expected to replace state regulation and ministerial decisions, granting more latitude to scientific institutions, to various degrees, to attain the agreed targets by largely autonomous ways. Therefore, the demands on the self-steering skills of the institutions and, consequently, on strategic control knowledge have increased considerably. At the same time, competitive differentiation means that university entrants, young scientists and collaboration partners of scientific institutions need more orientation knowledge. Not least, the increasing autonomy of scientific institutions entails that politicians and society demand more transparency regarding the performance of these institutions.

Against this background, public comparisons or rankings of scientific institutions, such as the university league tables published regularly by mass-circulation magazines, have become much more important. Their impact is far from limited to the eye-catching announcement of winners and losers. Rather, such rankings, which deal with the study conditions at various universities, aim to assist potential students with their decision where to study, and which subject, and thereby affect the recruiting prospects of individual scientific institutions. International ranking systems

[3] Wissenschaftsrat: Empfehlungen zur künftigen Rolle der Universitäten im Wissenschaftssystem. Cologne 2005.

– most notably the league tables of the Times Higher Education Supplement and the so-called Shanghai Ranking[4] – strongly influence not only the formulation of global science policy objectives, but also the strategic considerations of individual institutions.

Considering the consequences such rankings can have for scientific institutions, and their often uncritical reception, the lack of methodological transparency of many magazine ranking schemes, and the fact that the scientific community has no say in their development, must give cause for concern. Also, unfortunately for scientific research, the international ranking schemes fail to cover the non-university sector, which is of huge importance in many areas of scientific research, especially in Germany. Therefore in 2004, the German Council of Science and Humanities looked into the function and methods of comparative performance assessments in the sciences, drafted standards for such procedures, and presented a proposal for a procedure for assessing the research performance of universities and non-university research institutions.[5]

For its proposal, the Council started from the assumption that the existing rankings will stay and continue to exert considerable influence over the development of the science system. Therefore, the effects of such rankings can only be limited and contained by setting a differentiated, science-based procedure against the league tables produced by popular magazines. Still, unwelcomeand unintended effects even of the new procedure would have to be monitored and counter-balanced, if necessary, by differentiated communication of the results, continued development of the rating system and other accompanying measures.

The procedure proposed by the German Council of Science and Humanities provides for the assessment of the performance of universities and non-university research institutions in the dimensions research, promotion of young researchers, and knowledge transfer, whereas teaching is disregarded. The Council is well aware that teaching too is a fundamental task for universities and that the improvement of the quality of teaching is urgently desired. However, while the standards for the

---

4 The Times Higher Education Supplement: World University Rankings. 9. Nov. 2007. www.thes.co.uk; Institute of Higher Education, Shanghai Jiao Tong University: Academic Ranking of World Universities 2007. ed.sjtu.edu.cn/rank/2007/ranking2007.htm.

5 Wissenschaftsrat: Recommendations for rankings in the system of higher education and research. Part 1: Research. For an english translation see www.wissenschaftsrat.de.

comparative appraisals of the research performance of universities are internationally accepted and built on a solid theoretical foundation, this is not so for teaching. Apart from that, procedures assessing both research and teaching do not make sense, methodically. Consequently, no established, international example can be found for such an integrated procedure. Comparisons of teaching performances should be imbedded in a systematic initiative to improve the quality of teaching and learning, and they must be based on the efforts of individual universities to improve their quality management. Therefore, such comparisons will become reasonable only at a later stage.[6]

A.II. The pilot study on research rating

II.1. Previous history and decisions of the German Council of Science and Humanities

The Council presented the basic outlines of a subject-specific, multidimensional research rating system in its recommendations on rankings in the science system of 2004[7]. The procedure is based on the “Informed Peer Review” principle. In contrast to conventional ranking schemes, this does not involve the calculation of places in a ranking table. Instead, it means institutions are assessed by reviewers on the basis of standardized, statistically evaluated data about the individual institutions. The aim is to make competition within research more effective and efficient by increasing the transparency of research performance in the public sector. Another objective is to support the research institutions in enhancing their profile in the context of their respective mission by enabling them, through comparative performance assessments, to recognize their relative standing according to international established standards. Apart from that, the Council assumes that the results of research rating are of considerable interest to German and international, academic and non-academic cooperation partners as well as for all scientists and, especially, young researchers.

Having outlined the principles of a research rating procedure in its recommendations, the Council declared that, prior to a decision about the introduction of a regular process covering all disciplines, the suitability of the method must be tested through

---

6 The German Council of Science and Humanities is going to discuss recommendations concerning the quality of teaching at its spring meetings in 2008.

7 Wissenschaftsrat 2004

a pilot study. Therefore the Council established a steering group composed of members of its own Scientific Commission, other experts, and institutional representatives at vice president level from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), the Fraunhofer-Gesellschaft (FhG, Fraunhofer Society), the Helmholtz-Gemeinschaft (HGF, Helmholtz Association), the Hochschulrektorenkonferenz (HRK, German Rectors Conference), the Max-Planck-Gesellschaft (MPG, Max Planck Society) and the Leibniz-Gemeinschaft (WGL, Leibniz Association), as well as guest participants from German Länder ministries and the Bundesministerium für Bildung und Forschung (BMBF, German Federal Ministry for Education and Research). This steering group was commissioned by the Council first with the preparation and then, in July 2005, with the implementation of the pilot study. The selection of chemistry as one of the disciplines for the pilot study was also supported by the Gesellschaft Deutscher Chemiker (GDCh, German Chemical Society) and the Verband der Chemischen Industrie (VCI, German Chemical Industry Association), which agreed to contribute financially to the pilot study. As the second discipline, the Council chose sociology, as a subject whose methods clearly contrast with those applied in chemistry. The pilot study was started in October 2005.

The Council received an interim report on the progress of the pilot study in May 2007. In December 2007, the steering group published the results of the pilot study in chemistry, followed by the results in sociology in April 2008.[8]

II.2. Organization and implementation of the pilot study

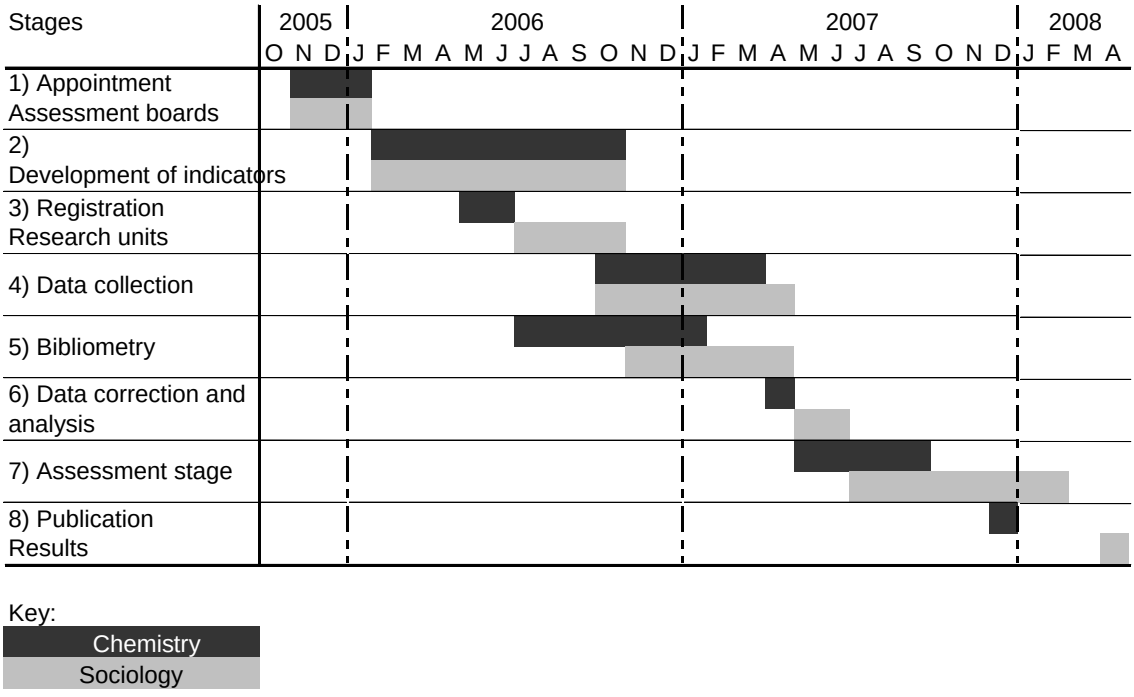
To adapt the procedure to the two disciplines selected for the pilot study and conduct the rating exercise, the steering group appointed two assessment boards composed of 15 or 16, respectively, expert reviewers. Candidates for the assessment boards were proposed by DFG, FhG, HGF, HRK, MPG and WGL and, in the case of chemistry, by the GDCh and the VCI. For sociology, the Deutsche Gesellschaft für Soziologie (DGS, German Sociological Association) was consulted, too. In its selection of assessment board members, the steering group was careful to ensure broad coverage of the principal areas of the respective discipline, and to win experts

[8] Steuerungsgruppe der Pilotstudie Forschungsrating im Auftrag des Wissenschaftsrates: Forschungsleistungen deutscher Universitäten und außeruniversitärer Einrichtungen in der Chemie. Cologne, 18.12.2007; id.: Forschungsleistungen deutscher Universitäten und außeruniversitärer Einrichtungen in der Soziologie. Cologne, 10.04.2008. For an English translation of summarized results in both disciplines see www.wissenschaftsrat.de.

with international experience. The international perspective was further formalized by including experts from the Netherlands, Austria and Switzerland. For both boards, practice-based representatives from organizations close to the respective discipline were appointed along with academic reviewers. In the case of chemistry, these practitioners are scientist from industrial corporations actively involved in research; for sociology, they come from a polling institute and from a charitable foundation. Finally, one rapporteur representing the steering group was delegated to each of the assessment boards and a project team to take care of the pilot study was installed at Council Head Office.

The main stages of the project study, and the periods they required, are shown in the diagram below.

Fig. 1: Implementation and schedule of the pilot study Research Rating



The first task of the assessment boards (Stage 2) was to determine the appropriate quantitative and qualitative assessment indicators for their respective discipline. Following a suggestion by the steering group, the Council decided to adopt as a general condition that the quality of research should be the core criterion of research rating and be assessed in a more differentiated way than the other criteria. To this end, so-called “research units” were defined, generally below the faculty level at

universities, or at the departmental level at non-university institutions. The aim of this stage was to assign subject-specific indicators to certain criteria, forming a so-called assessment matrix. The first draft of this matrix was trialed in a pretest and corrected and simplified accordingly. The final, published assessment matrix was used as the basis for the remaining data collection and assessment process.[9] In consultation with the steering group, the number of criteria was reduced from nine, as originally suggested by the Council, to six in order to make the procedure more manageable.

Fig 2: Table of dimensions and criteria after streamlining by the assessment boards

Dimension	**Criteria**
Research	I. Research quality (at research unit level)
	II. Impact/Effectiveness
	III. Efficiency
Promotion of young researchers	IV. Promotion of young researchers
Knowledge transfer	V. Transfer to other areas of society
	VI. Promotion of the public understanding of science

This stage also included the task of defining more precisely the different areas of the respective discipline, to define the grades of the rating scale, and to develop the questionnaires used for data collection.

In preparation of data collection, the research units to be assessed and the scientists associated with them – in chemistry: the senior scientists and group leaders – at the participating universities and non-university research institutions had to be registered (Stage 3). Therefore the institutions were asked, initially, to nominate one scientist for each of the two disciplines to act as the so-called subject coordinator. Due to the tight schedule of the pilot study, the nomination of the subject coordinators and the registration of the research units were conducted parallel to the development of the indicators by the board members.

[9] For details of the assessment matrices and specific indicators, see the final reports of the two assessment boards, available only in german; for the Pilot study Research Rating Chemistry: Abschlussbericht der Bewertungsgruppe (Wissenschaftsrat, Drs. 8370-08); for the pilot study Research Rating Sociology: Abschlussbericht der Bewertungsgruppe (Wissenschaftsrat, Drs. 8422-08).

For the data collection stage (Stage 4) questionnaires and preformatted tables, which could be e-mailed and edited using commercial office software, were sent to the subject coordinators. Depending on the institution and the discipline, a general questionnaire and one further questionnaire per research unit, including tables, had to be filled out. Since representative citation data are not available for sociology, the research units were also asked to submit a certain number of exemplary publications in electronic format. A five years' survey period, 2001-01-01 to 2005-12-31, was defined for all data and other information.

Once the research units were registered and their scientists nominated – for chemistry this stage had been completed even before the questionnaire forms were finalized –, the Institut für Wissenschafts- und Technikforschung (IWT, Institute for Science and Technology Studies) at the University of Bielefeld and the Informationszentrum Sozialwissenschaften (IZ) of GESIS (German Social Science Infrastructure Services e.V.) in Bonn were assigned for chemistry and sociology, respectively, to compile the publications of the registered scientists from the survey period. The resulting lists were to be used for publication analysis and, in the case of chemistry, citation analysis (Stage 5). For chemistry, Thomson Scientific's Web of Science was used for this purpose. For sociology, the IZ used a combination of its own SOLIS database and various databases of Cambridge Scientific Abstracts. Both institutions put the results of their searches on password-protected Internet sites and asked the institutions to be assessed to check and, if necessary, amend the respective lists of publications.

Council Head Office undertook the task to check all data for compliance with the collection rules and for consistency and plausibility and, in consultation with the subject coordinators, applied the necessary amendments (Stage 6). After that, Head Office aggregated the results of the survey with other data obtained from external cooperation partners into one data report per institution and discipline. These reports were returned to the respective institutions for final checks, before Head Office undertook the statistical analysis. In this, the calculation of percentiles[10] for all quantitative indicators was of central importance. These values enabled the

[10] The percentile is a value indicating how many of all units assessed achieve no better than the unit considered, with regard to a certain indicator. For instance, a percentile of 90 % regarding the number of publications by a certain institution means that 90 % of all institutions have registered fewer or, at most, the same number of publications as that institution. The percentile always ranges between 0 % (lowest value observed) and 100 % (highest value). The percentile of 50 % is also referred to as the median.

reviewers to quickly assess the relative standing of the respective institution in their discipline, independent of any specific measuring unit.

Based on the data reports, the lists of publications of the research units and, for sociology, the actual publications submitted to the assessment board, the experts rated the institutions and research units according to the individual criteria (Stage 7). Initially, two experts per case were assigned as rapporteurs, of which each was asked to propose their rating, independently. If there was disagreement between the ratings of the two rapporteurs, they were given the opportunity to clarify the reasons for this before all ratings underwent plenary discussion by the assessment boards and, finally, moderation in the context of the overall distribution of the results.

II.3. Experiences from data collection[11]

In response to the inquiry sent out in April 2006 to all state-owned universities and selected non-state universities and to the organizations operating non-university research institutions, if they would take part in a research rating exercise for chemistry and/or sociology, 78 institutions for chemistry and 64 institutions for sociology came back with a positive answer. Since a number of institutions withdrew for various reasons (see below, p. 17) while the study was underway, the exercise was completed with 57 universities and 20 non-university institutions for chemistry and 54 universities and three non-university institutions for sociology. The research units registered by the reference date, 2005-12-31, included 1038 chairs of chemistry and 376 chairs of sociology. For chemistry, this figure is significantly higher than the number of university professors teaching and researching in chemistry quoted by the Statistisches Bundesamt (895 in 2005[12]). For sociology with its diverse subject classification, there are no official statistics to provide comparative figures. However, a comparison with the comprehensive survey of the German Sociological Association from 2004 justifies the assumption that, apart from above-mentioned withdrawals, the corresponding survey for the pilot study was comprehensive, too.

In some cases data collection put considerable strains on the subject coordinators, because the data, most often at universities, were not available centrally, but only at individual institutes or even from individual scientists. At some institutions, support

[11] The experiences from the pilot study, outlined in this and the following sections, were taken from the final reports of the assessment boards, where they are described and evaluated in more detail.

[12] Statistisches Bundesamt (edts.): Fachserie 11, Reihe 4.4, Personal an Hochschulen, Wiesbaden 2006.

from administration, for instance by supplying data on third-party funds, did not come forward as smoothly as hoped for. Also, many subject coordinators found the handling of the electronic questionnaires, with their preformatting and text length limitations to ensure their uniformity and processability, too laborious. Many would have preferred an actual online survey at that stage.

The defined survey period was 2001 – 2005 with 2005-12-31 as the deadline date. For chemistry, the so-called “Work Done At“ principle was applied, meaning all scientists working at the respective institution within the survey period were assigned to that institution with their contributions achieved there during the period, even if they had moved to another institution or retired before the end of the period. In sociology, on the other hand, the “Current Potential” principle was followed, according to which only the scientists employed on the deadline date were assigned to the respective institution, with all their contributions delivered during the survey period.

The “Current Potential” principle entails that any posts not occupied on the deadline date were left out of the rating exercise. On the one hand, this allowed a more up-to-date performance assessment, also with regard to achievements expected in the future; on the other, however, it contributed to the outcome that in sociology the fraction of research units that could not be rated was significantly higher, at 7 %, than in chemistry (2 %). It was also among the reasons for the withdrawal from the study, following the announcement of the survey terms, of seven universities in sociology compared to only one university in chemistry.

The research units, which were to allow research quality assessments differentiated at a lower level than that of entire institutions, were defined differently by the two assessment boards, in line with the different cooperation practices in their respective discipline. As a result, the units registered in chemistry were generally larger – with six senior scientists, on average, including three professors – than in sociology, where nearly 75 % of all units comprised only one professorial chair. Sometimes the subject coordinators found it difficult to assign interdisciplinary units. In chemistry these were in the minority, but in sociology the majority of the institutions consider themselves engaged in interdisciplinary research. Still, interdisciplinarity was seen as an issue more often in chemistry, with particular concerns about the “artificial” nature

of the research units to be formed and on the difficulty of comparing interdisciplinary units with those at the core of the respective field.

The option to register transinstitutional research units, which was intended for cases of joint appointments, was used in only few cases. Other subject-specific peculiarities of the research unit structure (service units, affiliated institutes) are discussed in the final reports of the assessment boards.

Due to the tight schedule of the pilot study, the end of the survey period had already passed when the questionnaires were sent out. The consequent, retrospective nature of the survey presented considerable difficulties for the many universities that had never been confronted with comparable data requests. This issue was often cited as the reason why the originally intended data collection period of two months was too short. Therefore, following requests from many subject coordinators, the period was extended by six weeks. For the future, to allow concurrent data collection, many subject coordinators suggested to announce the precise data requirements with a substantial lead time ahead of the survey deadline.

For 85 – 90 % of all institutions in both disciplines the data inspection at Head Office indicated that amendments were necessary, requiring further consultation with the respective subject coordinator. This need for amendments, the subject coordinators' appraisals of the costs of their data collection work, and the value of the data for the reviewers are subject of the recommendations for improving the data basis in final reports of the two assessment boards.

There are clear differences between the two disciplines of the pilot study regarding the data basis available for bibliometric analysis. The databases from the Thomson Scientific Web of Science, which were used for chemistry, comprehensively cover the periodical literature in this discipline and allow a citation analysis. There were no objections from within the discipline against using these databases. Since types of publication other than journal contributions are not covered by these databases, the survey also allowed manually entering other publications in the data collection forms. The volume of such additions – mainly contributions to compendia, encyclopedias and conference proceedings – accounted for about 5 % of the periodical contributions recorded in the Web of Science.

In contrast to that, the coverage of publications by the participating sociologists in the databases SOLIS and CSA, which do not support citation analysis, is rather sketchy. The lists of publications originally found in the databases roughly doubled in length in the course of revision. Since not every scientist took part in the revision, the existing databases are probably even less complete than indicated by the pilot study. About one quarter of the additional entries were international publications, while three quarters were German publication not recorded by the German social sciences literature database, SOLIS. One central reason for the incompleteness of the databases is the highly diversified publishing practices of sociologists. In sharp contrast to the situation in chemistry, contributions to compilation volumes are most common among sociologists, with a share of about 45 % of all registered publications, followed by essays published in journals (34 %, about a quarter of which is recorded in international citation databases) and monographs (7 %). Also, the definition of sociology as a discipline is more blurred than is the case for chemistry. As the majority of the sociology research units describe themselves as interdisciplinary, a large fraction of their publications appears not in periodicals that can be clearly identified as sociological journals, but in a broad range of more than a thousand periodicals with widely varying specialist subjects[13] and in compilations. To a considerable extent, sociologists' writings are handled by publishers whose quality assurance system for scientific publications is uncertain. In a discipline with such publication practices, the body of publications cannot be recorded completely relying on a small number of databases. Therefore, a comprehensive survey can only succeed if the scientists or research institutions concerned cooperate in the necessary amending of the data pool, which has to be organized by a central service facility.

II.4. Experiences from the assessment process

Dividing the assessment stage into two phases, first the independent preparation by two rapporteurs per unit and then the plenary discussion of all ratings, concluded by a consistency check, has proven to be a very successful approach. Taking into account their specialities and possible biases, the large majority of the research units could be covered by rapporteurs from the respective assessment board. For

[13] While the approx. 40,000 items of periodical literature in chemistry were spread over about 1,700 journals, the much smaller number of sociology articles, about 4,000, appeared in more than 1,000 different periodicals. To cover 50 % of articles, 76 sociological journals had to be evaluated, whereas only 47 journals needed to be surveyed to achieve the same coverage in chemistry.

sociology, research units employing a member of the assessment board were generally rated by an external, special reviewer and the chair of the assessment board. In chemistry, the initial assessment was usually undertaken by an external, special reviewer and a member of the assessment board. Some very large research units were assessed by three rapporteurs. In just 8 % of all cases special reviewers were consulted in order to exclude biases, rate highly specialized research units, or call in additional opinions on cases where a disagreement between the rapporteurs could not be resolved by themselves. To avoid overstretching the relevant experts on the assessment board, special reviewers were also employed in a small number of cases from very large branches of the respective disciplines.

The individual assessment phase took about six weeks, in chemistry, and twelve weeks in sociology. While the longer term required in sociology had been planned for from the beginning, because of the large number of exemplary publications that had to be read, it also proved necessary because the relatively poor quality of the publication data called for more additional research.

To facilitate the plenary moderation, the rapporteurs were informed in advance about the cases in which their rating differed from that of their respective co-rapporteur. Overall, there was a high level of agreement between the independent assessments by these experts: Depending on the criterion, both rapporteurs proposed identical grades in 75 – 86 % (chemistry) or 71 – 89 % (sociology) of all cases. The final results are independent of who were the rapporteurs assigned to a unit. Any reviewer's bias potentially affecting individual cases, and any predeterminations were neutralized by detailed discussions within the heterogeneous assessment boards, while possible tendencies towards conformity were obviated through critical discussions and by experts taking turns in playing the role of devil's advocate in the plenary sessions. It also appeared important to avoid undue pressure to arrive at any decision by allowing to describe individual units as "unrateable".[14]

For any procedure based on empirical methods, such as research rating, errors are impossible to avoid completely. Based on the existing data, the expert votes and their quality inspection in the plenum, the assessment boards conclude that the likelihood of incorrect ratings is relatively low and the extent of such incorrect ratings, should

[14] On possible mistakes committed under pressure to decide, see Janis, Irving. L. (1982). Groupthink. Psychological studies of policy decisions and fiascoes. Boston: Houghton Mifflin.

they occur, does not exceed a single grade level. However, considering that research units were classified as “unrateable” in cases of uncertain data or inextricable disagreement, incorrect ratings can only have occurred in very few cases.

In chemistry the reviewers decided to add another level of differentiation in their assessment of the research quality, for which the best data were available, by introducing an additional grade level, “very good to excellent” between the grades “very good” and “excellent”. The criteria within the knowledge transfer dimension, on the other hand, appeared more problematic, especially the criterion “Promotion of the public understanding of science”. Due to the mainly qualitative and very heterogeneous data basis, the assessment board for chemistry decided to apply only three different grades to the latter criterion, “below average”, “average” and “above average”. The assessment board for sociology followed this decision and used the simplified grade scale for the criterion “Transfer to other areas of society”, too. These changes to the ongoing procedure necessitated reassessments in some cases.

At the stage of the plenary consultations it emerged that in some special cases[15] it was impossible to arrive at a rating. To briefly give reason for the “unrateable” vote in such cases, but also to make transparent the reasons for a certain rating or to qualify the rating, the two assessment boards employed the instrument of commenting individual assessment results.

II.5. Summary and reception of the results

The results of the assessments were adopted and published by the steering group.[16] The assessments according to the six criteria of the research rating procedure for every institution taking part from the respective discipline were presented in two publications. Additionally, each of the publications contained an anonymized profile representing the grade distribution of the weighted research units of the individual institutions for the “research quality” criterion. For reasons of data protection, the results for individual research units were not published, but only communicated to the respective institutions and the state ministries or operating organizations in charge of them.

---

[15] Depending on the criterion, 2 – 6 % of cases in chemistry, 4 – 7 % of cases in sociology

[16] See above, fn. 8, p. 12.

The full range of grades was used in both disciplines and for all criteria. In chemistry, the ratings for every criterion were distributed symmetrically about the grade “good”, which also is the most common result, with a share of 25 – 40 % of all assessments (depending on the criterion). The top grade, “excellent“, was awarded to 5 % of all units[17] for the central criterion, “research quality”, and in 6 – 12 % of all cases according to the other criteria assessed at institution level. In sociology, the average grade for research quality was just “good”, with 4 % of the research units awarded an “excellent” grade. At institution level, 7 – 9 % of them were rated as “excellent“. The assessment of the processes for the promotion of young researchers in sociology is significantly less positive than for chemistry, where this criterion counts among the strength of institutions in Germany.

Some of the criteria, especially in the research dimension are interconnected in their content. This, however, does not make any of them redundant. Institutions rated as about mid-table in terms of the impact/effectiveness criterion can be excellent with regard to research quality, efficiency, promotion of young researchers or knowledge transfer whereas, vice versa, institutions with a very high research profile are not always the most efficient or high achievers in terms of the promotion of young researchers or knowledge transfer. The great importance of multidimensionality for the acceptance of research rating is manifested by the institutions’ reactions to the publication of the results, in which they often focus on one criterion that is central to the profile of the respective institution.

The assessment of research quality even of the research units within individual institutions produced much differentiated results. In more than half of all cases both in chemistry and in sociology, the ratings for research units within a single institution are spread across at least three grades. It became obvious in the course of the pilot study for chemistry that there could be considerable interest in the publication of the results not only for entire institutions, but also for individual research units. As mentioned above, while such results were disclosed to the participating institutions only for internal use, the have been published so far only in anonymized format. Publication of the results at research unit level would make it easier for the institutions to interpret them in comparison with the respective direct competitors of individual research units. Moreover, this information would be very useful for external

[17] in addition to 7 % “very good to excellent“

users, e.g. potential cooperation partners and young researchers. One objection against the publication of the results achieved by research units is that some of the units are composed of less than three (senior) scientists, a fact that could turn the ratings into personal data, which would be subject to data protection regulations. Concerns regarding data protection could be avoided by effectively precluding personal identification of participating scientists or by securing the advance consent of the scientists to publishing the results.

Considering the present results, one could ask if the elaborate assessment procedure applied in the pilot study could be replaced by an optimized ranking system, in which an overall rating is computed from weighted, quantitative indicators. This option would be conceivable, in principle, for criteria I to IV[18]. According to the best models tested so far, quantitative ranking would produce results differing from the present reviewers' assessments by at least one grade level in between 20 and 30 % of all cases, depending on the criterion. For "research quality", this figure would even rise to 36 %. This question should be studied by further analysis of the benefits of a research rating system, despite the fundamental drawbacks of quantitative ranking compared to an informed peer review, even if the ranking produced a good prognosis of the review results (see section B.I, p. 27; for the comparability of the results with previously published rankings, see annex, p. 46 ff).

Already, there have been comprehensive press reports about the results of the pilot study in chemistry and the new procedure. The majority of reports highlighted the distinguishing characteristics of research rating compared to conventional ranking schemes, especially the principle of the informed peer review and the multidimensionality of the assessment. Several writers welcomed the fact that the Council was setting new standards for performance comparisons in science with its new procedure. So far, there have been no known attempts to translate the results of the pilot study into headline-grabbing league tables. However, it should be noted that the different grade scales used for the knowledge transfer criteria presented a complication that caused some media reporters to focus exclusively on the assessments of the research dimension.

[18] Research quality, Impact/Effectiveness, Efficiency, and Promotion of young researchers. For the criteria "Transfer to other areas of society" and "Promotion of the public understanding of science", with their lack of quantitative data, this would be out of the question.

After completion of the pilot study, the assessed institutions also received, apart from the published and confidential rating results, the data reports containing the statistic evaluations of the data submitted by them, and an overview paper outlining the survey mode and the national distributions of the quantitative data (the same overview that the reviewers had used to arrive at their ratings). The detailed inquiries received about these materials indicated that the survey and the analysis and summarization by Council Head Office generates additional benefits for the assessed institutions.

II.6. Costs

The costs of the pilot study consist of three components, which can be quantified with varying precision:

- The direct costs for the administration of the procedure: costs for the organization of meetings, travel costs and expenses payable to the reviewers, and costs for the publication and citation analyses, including license fees. These direct costs amounted to approx. € 1.1 mil. for chemistry[19] and sociology together over the entire duration of the pilot study.
- The working hours of the 15 or 16 reviewers per discipline, respectively, which amounted, according to the reviewers' estimates, to 4 – 5 working weeks per person in chemistry and 8 – 10 working weeks in sociology, accumulated over the duration of the pilot study. The main reasons for the difference between the two disciplines in this respect were: The assessments in sociology required wider reading of selected publications; and the poorer quality of the data necessitated more intensive work with the raw data.
- The indirect costs incurred by the participating institutions for data collection and amending of lists of publications. These costs are impossible to quantify with any precision, because the figures supplied by the institutions varied widely. According to a worst-case estimate, the cost per institution could have amounted to up to two person-months, with wide differences between institutions, and there is no information about the relative share of scientists' and administrators' work invested.

[19] The Chemical Industry Fund contributed approx. 25 % of the prorated costs for the pilot study in chemistry.

These estimates refer to the first research rating exercise ever performed in Germany and include both disciplines. Factors that would influence the costs of research rating if the procedure were expanded or repeated are discussed in section B.II.6, p. 43 ff.

B. Recommendations

B.I. Recommendations on the future of research rating

Research rating offers a range of unique characteristics distinguishing it from common ranking schemes:

- The quality of research is differentiated and assessed by an informed peer review based on quantitative and qualitative comparative data and taking into account context information.
- The relevant learned societies in the respective disciplines contribute to the definition and operationalization of the assessment criteria.
- The documented differentiation of the quality of research within the respective institutions enhances the informative value of the results.
- The rating according to a range of criteria ensures that the results reflect the variety of performance profiles of different research institutions.
- By including non-university research institutions, which play an important role in disciplines such as chemistry, the rating provides a comprehensive picture of Germany's research landscape.

The pilot study for one natural science and one social science showed that the research rating procedure is feasible and produces meaningful results. As explained in the reports of the assessment boards, the different rating criteria are of varying robustness. In the view of the steering group, the pilot study is part of a learning process towards the gradual, further development of a research rating system. This development should include further clarification, in dialog with the users, how the procedure could produce optimum benefit and, at the same time, its costs could be limited.

In the view of the steering group, the research rating system developed by the German Council of Science and Humanities can, due to its unique characteristics, perform a number of functions that can not be fulfilled satisfactorily by existing procedures:

- Research rating provides impulses for scientific institutions to develop their own strategies, and it can serve as a success control mechanism.[20] It supports the institutions in creating a successful profile within a science system that is organized, increasingly, along competitive lines. To achieve this, their management bodies need reliable assessments of their strengths and weaknesses in comparison with their direct competitors.[21]
- Research rating produces data about the quality of actual scientific performance that is significantly more valid than information e.g. from the evaluation of third-party funding statistics and other quantitative indicators in conventional rankings. Thereby, it meets the demand from science politics for reliable data about the performance of scientific institutions – a demand that is going to grow to the extent to which science funds will be shared out increasingly through competition.
- By representing the performance of German scientific institutions transparently and in a format comparable with internationally established rating systems, research rating raises the international visibility of German science. In the same way it provides a counterbalance to international rankings that do not recognize the non-university research landscape in Germany.
- By subject-specific assessments, differentiation of criteria, including an efficiency rating related to the use of personnel, and especially by assessing individual research units the research rating procedure allows reliable identification of good research performances, including those achieved outside the leading institutions in Germany. Considering the present focus on international excellence, it is still important to improve the awareness of promising activities especially at minor research locations, in order to maintain the broad basis of scientific quality on which top achievements are founded.
- By supporting the research rating process, and taking it as an opportunity for critical self-reflection, the science system is contributing to improved transparency and more efficient use of funds. In this way research rating helps the science system to deliver the accountability demanded by society and politics.
- Research rating emphasizes the promotion of young researchers as an important task, in its own right, of scientific institutions. The research ratings help young

---

[20] Research rating does not provide explanations for the high or low level of performance of any individual institution. Neither can it deliver an ex-ante assessment of new strategies. Therefore, it replaces neither detailed individual evaluations nor the work of advisory committees.

[21] For comparisons it is important to have access to as much detailed information as possible about competitors' ratings. This makes publishing the results for individual research units a favorable prospect, provided the data protection and legal issues involved in such publication can be resolved (cf. B.II.5, p.41 f.).

researchers and advanced students from Germany and abroad to form a judgment about which institutions would offer a suitable environment for the first steps of their scientific careers.

- In disciplines such as chemistry, research rating provides an important orientation tool for potential cooperation partners from the corporate sector where, in the age of globalization, the choice of location increasingly depends not least on the attractiveness of the research environment.

In contrast to existing rankings, the research rating procedure offers the advantage of providing quality assessments that are more differentiated and more reliable than ranking tables based on quantitative data only. An analysis of the results shows that in more than a third of all cases the core criterion, research quality, in particular would be rated differently if the grades were computed simply by weighting the quantitative indicators, without further reviewing. Even deviations of only 20 – 30 %, as found for the other criteria, are not acceptable, showing that the function of the reviewers in the informed peer review is indispensable. On the other hand, without the comprehensive data collected for the research rating even experienced reviewers would be in no position to differentiate reliably between research units apart from those known as particular active units. Also, in some cases, the appraisal of the data made the reviewers assess well-known units more critically than would have been expected on the basis of the units' reputation. Thus, the added value of the research rating arises from the combination of the expertise and experience of the assessment board and a comprehensive data basis suitable for statistical evaluation, but also comprising some qualitative components.

The fact that research rating is a review procedure must not be misinterpreted in the sense that it were a collection of schematic evaluations of individual institutions, or could replace such evaluations. Rather, the procedure is valuable because it compares a multitude of institutions according to uniform standards, whereas it is not intended to analyze the reasons for the performance level of any individual unit and to produce recommendations for its improvement.

Should research rating be continued, one had to be careful that procedures aiming to strengthen performance competition within the science system would also bring about unintended incentive effects, apart from the intended ones. Such effects

should be continuously monitored and critically discussed. Even if there is no algorithm for a direct conversion of indicators into ranking positions, just by the fact that certain data are collected at all, these data become significant enough that the assessed institutions will consider it a rational decision to invest resources into optimizing the processes measured by the data in question. The approach followed by the research rating procedure – to assess several performance dimensions side by side, without weighting, and to base the rating not only on quantitative data, but also on qualitative information obtained by open questions – serves to limit such homogenizing effects as far as possible. Too much streamlining of the procedure would destroy this crucial advantage, not least in terms of its acceptance within science.

The experiences available so far indicate that research rating, with appropriate adjustments to the procedure, can be transferred as intended to other science areas. Major modifications must be expected for the rating of disciplines from the arts and the technical sciences, since those are the fields most different from natural and social sciences, both in their internal publication and communication channels and in their relations to other areas of society, which are relevant for the transfer dimension. Accordingly, such disciplines should be considered next, if the procedure were to be developed further.

Further development of research rating should involve dialog with the users of the ratings in order to get a clearer picture of how the published results are actually applied, not least in comparison to any internal evaluations that might exist. After that, the recommendation of the assessment board for chemistry, to re-assess chemistry research after an interval of five years and to assess biology and physics at the same time, should be considered again.

One essential advantage of the research rating carried out for the pilot study is that it was designed with input from the science system. For the medium term, research rating will be acceptable only if possible, unintended consequences of the incentive effects of such rating exercises are be taken into account in the development of the procedure. Therefore, the steering group recommends to the Council to support the examination of the effects of research rating and related rating procedures.

B.II. Optimization of the procedure for a research rating

II.1. Organization and implementation

The size and composition of the assessment boards chosen for the pilot study proved to be reasonable. While for minor disciplines somewhat smaller boards could be conceivable, larger boards are not recommended. The international experience of the reviewers is particularly important to ensure that the assessment conforms to international standards, especially in disciplines where comparable international data are not available. The proportion of female reviewers should be increased. The recourse to candidate suggestions by the scientific organizations and the involvement of the learned societies in the recruitment of reviewers not only brought pragmatic benefits, but was also essential in securing the acceptance of the procedure within the subject communities. This approach should be continued under all circumstances. Generally, the reputation of the reviewers within their discipline is very important for the acceptance of the procedure.

According to statements from the assessment boards, the comprehensive support by Council Head Office was of great importance for the success of the research rating process and, if the system is going to be continued, must be secured for the future so that the workload for individual reviewers can be kept within acceptable limits. This is particularly important should the rating be repeated for the same discipline, because in that case the special incentive of taking part in a pioneering experiment would be absent. The willingness of renowned scientists to serve as reviewers in a research rating exercise crucially depends on them being assured that they will receive appropriate support and can rely on a clean data basis. Head Office can undertake the conceptual preparations, which were done for the pilot study and would be necessary for disciplines that are rated in this way for the first time, only if one research associate can be assigned to each discipline. For repeat ratings of the same discipline, these preparations would be expected to become more routinized and require less personnel input.

The way the pilot study was implemented was generally successful and should be maintained for future research rating exercises. For each discipline to undergo a rating procedure of this type it is important to provide sufficient time for developing the indicators, so that newly developed indicators can be pre-tested, if necessary.

For the assessed institutions it was of concern that they were asked to form research units before they knew exactly how the research performance of the units would be assessed. Such errors of sequence should be avoided in the future. Another criticism concerned the shortness of the lead time for data collection (see II.3, p. 35 f.). By allowing more time for the institutions to prepare, it will be possible to conduct the actual data collection within a fixed time window immediately after the deadline, with additional days of grace only in exceptional cases. This will also have the effect that the data are more up-to-date at the time of assessment.

II.2. Subject of the assessment

a) Subject areas and interdisciplinarity

For a comparative assessment of research performances it is essential to define appropriate comparison groups. For the pilot study, the comparison groups were defined according to disciplines or subjects. The main point in favor of this approach is that quality standards in science are determined primarily within the scientific subject communities. The disadvantage is, however, that institutions that are active in several disciplines need to be subdivided into subject areas and the interactions between the disciplines is not considered in the assessment. More detailed differentiation between disciplines leads to more frequent difficulties with the classification of research activities under one or the other subject area. On the other hand, an assessment based on a system of disciplines that distinguishes too roughly between only a few different subject areas can be distorted, because uniform standards would be applied to very diverse research practices. This balance needs to be monitored, should research rating be continued.

The conflicts of classification encountered in some cases in the course of the pilot study were mostly attributable to the incompleteness of the data basis: Scientists were not sure about possible alternatives to the classification under chemistry or sociology, respectively. Should the procedure be established as a permanent system, these uncertainties could be avoided by publishing, prior to collecting any assessment data, a taxonomy of all subject area to be considered for rating.[22] Such taxonomy should be based on the system of currently 48 sectional DFG Review Boards with their respective subject areas. Care must be taken that the subject areas

[22] See Wissenschaftsrat 2004, p. 44.

are not defined in too great detail, since for subject areas much smaller than sociology,[23] the costs/benefits ratio of research rating would turn unfavorable. In such cases, if possible, adjacent subjects should be subsumed under a common subject area. This would also reduce the difficulties of subject classification and interdisciplinarity. Conversely, the workload for the reviewers sets an upper limit to the size of those subject areas that, due to the absence of citation data, can only be assessed by reading of selected publications. The upper limit of 50 subject areas, as advised by the Council in 2004, still appears realistic.

Also, when defining the subjects, the typical and predominant definitions of organizational units at universities must be taken into account, since data collection across different organizational units entails higher costs for the institutions, and the results would be less relevant for steering purposes. To find a compromise that would be acceptable for all concerned, a draft taxonomy should be discussed through a public consultation process with universities, scientific organizations and learned societies before a research rating procedure for several disciplines is based on it. This process could be commenced for instance in the course of the preparations for the possible extension of research rating to the natural sciences biology, chemistry and physics.

If possible, interdisciplinary research activities should be assessed as coherent units without artificial divisions. To obviate possible difficulties in their assessment, scientists from the fringes of the respective discipline should be among the reviewers appointed to the assessment boards. Also, the assessment boards in the pilot study found it a positive experience to be able to consult special reviewers for highly specialized and interdisciplinary research units.

b) Research units

For the definition of the research units, the two disciplines assessed in the pilot study used different approaches. Registering the names of the respective scientists has proven a reasonable starting point also for the subsequent publications search. Because of the connection to the publications search, the personal registration should not be restricted to certain categories of personnel, as was the case, initially,

[23] Sociology does not have its own DFG Review Board; it is one of the subject areas allocated to the “Social Sciences“ Review Board.

in chemistry (professors and group leaders) at this stage, but include every independently publishing scientist. The differences between the disciplines mainly concern the formation of the research units.

Considering the degree of detail of the research units defined in sociology, the assessment board for sociology discussed the proposition to stipulate that in future – similar to the decision of the assessment board for chemistry – only entities at institute level (Institute for Sociology, Institute for Medical Sociology, etc.), but not at the level of individual, professorial chairs should be registered as research units. This measure would reduce the number of research units in sociology by more than half. Moreover, such larger units are subjects of strategic steering, which the research rating was intended to support, and can be strengthened in their qualities as active stakeholders. However, the assessment board arrived at the conclusion that the detailed structure registered in the pilot study properly reflects the current structure of the discipline. Although the development of larger units in sociology, characterized by common research programs, shared infrastructures and increased continuity, would be desirable, it is the opinion of the assessment board that their definition as fictitious entities, just for the purposes of research rating, would diminish the informative value of the results to an unacceptable extent, not least in view of the often significant variations of performance.

With a view to ensure acceptance within the science system, it appears reasonable to maintain some leeway for setting out different conditions for the research units, as appropriate for the respective cooperation practices in different fields of science. However, due to the experimental nature of the pilot study and the uncertainties entailed by this, the variance between participating institutions within the individual disciplines was wider than absolutely necessary. Improved standardization should be ensured by early consultation and agreement with the reviewers about the final breakdown of institutions into research units.

c) Non-university research institutions

The inclusion of non-university research institutions in the research rating exercise is one of the great advantages of this procedure over existing national and international ranking schemes. In many subject areas, including chemistry but not sociology, non-university institutions contribute an important share of the volume and quality of

German research achievements. The best non-university institutes can even serve as a benchmark for international research quality, and thus help to calibrate the assessment scale. Without the non-university institutions, the value of the picture of Germany's research landscape painted by a research rating would be much diminished.

Many non-university institutions are organized as multidisciplinary facilities and consider this form of organization a value in itself. The question how this can be taken into account in the rating procedure requires further examination.

Specific solutions are needed for research units whose head scientists simultaneously hold posts at a university and at a non-university institution. The offer extended to the eight participating institutions – to allow registering "shared units" of two partner institutions, for which an integrated data set would be produced, which would be assessed as one, and whose result would be accounted to each of the institutions at proportions to be agreed between them – was taken up in only one case. Conversely, some institutions appeared to have difficulties with the alternative chosen for the main survey, by which the respective head scientists were associated to both institutions, while the research achievements were clearly separated. In isolated cases, major data overlaps, which could not be explained in any detail, made the research units unrateable. If the research rating procedure was to be developed further, both alternatives – "shared research unit" and "separation" – should be offered and the consequences of each choice be clearly communicated in advance. In the medium term, based on the experiences from this options model, rules should be developed for deciding which option should be chosen in which cases.

II.3. Data collection and analysis

Any optimization of the data collection process must pursue two objectives: to further improve the data quality, so that it becomes easier for the reviewers to assess the data, and to reduce the workload for the institutions providing the data. Therefore, careful preparation of the data collection process for each discipline by the assessment board and Head Office, including possible pre-tests, would be vital if research rating were continued. In the medium term, the burden on the institutions should also be limited by improved consultation about collection formats between institutions collecting research data, thus enabling multiple use of the data.

The development of indicators leads to the compilation of an assessment matrix for each discipline, in which the individual criteria are defined in more detail by so-called assessment aspects, and assigned to the indicators. This matrix provides the structure for the subsequent data collection and analysis stages.

Data collection in the individual institutions should remain the responsibility of a subject coordinator nominated by the institution. The proven approach of the pilot study was to appoint a subject scientist for this task. It also proved important that the subject coordinator received appropriate support from the administration of the respective institution. Institutions should be advised against delegating the data collection task to junior researchers, who most likely lack solid information about the focus areas of the research units of their discipline within the survey period and are probably not aware of the data sources existing at their institutions.

To be able to fulfill their task, the subject coordinators must receive thorough and reliable information about every detail of the procedure well in time. Therefore, data collection should begin only after publication of the assessment matrix. In this context it is worth noting that, from the perspective of the assessed institutions, the registration of the research units is already part of data collection. Information events held prior to data collecting would be helpful. The extended lead time requested by many subject coordinators – several of them suggested to follow the example of the Research Assessment Exercise and publish the assessment matrix and the questionnaires based on it ahead of the survey period, to allow concurrent data collection – would contribute to speedy implementation of the data collection stage. Apart from reducing costs, this would offer the advantage that the data would be more up-to-date at the time of their assessment by the reviewers than in the pilot study. Therefore, if a rolling system was to be introduced, step by step covering all subject groups, a gradual extension of the lead time for the data collection stage should be envisaged. In the same context, the data collection formats should be further standardized to allow multiple use of the data and enable the institutions to operate a data retention system for a range of purposes.

For the pilot study, publication data were not collected directly from the institutions. Instead, existing publication databases were searched and the publication lists from the searches amended in consultation with the subject coordinators. This procedure

is well established for disciplines with internationalized publication practices, such as chemistry, and produces good, reliable results met with a high degree of acceptance. In contrast, the publications search in the existing databases for sociology was much less successful. Apart from the low level of internationalization in sociology – 85 % of the publications registered in the end had been published in the German-speaking area – and the more blurred definition of the discipline, this is also connected to the fact that articles published in periodicals account for only about a third of sociological literature, and monographs and compilations are more difficult to record systematically. This finding would also apply to other subjects from the arts and the social sciences whose publication culture is largely national and, in terms of the spectrum of publishing organs, highly diverse. Consequently, extending the research rating system to more of those subjects will entail the necessity to amend and, to a considerable extent, re-register literature data in cooperation with the respective scientists.

For cost reasons it would also be conceivable in those disciplines to apply a selective approach, e.g. restricting the publication lists to periodicals recorded in certain databases. In that case, however, care must be taken that the incentive effects caused by this approach do not result in unwelcome changes of the communication structure of the scientific discipline in question, and thus to disadvantaging of interdisciplinary or practice-based research. Against this background, a comprehensive registration system capturing all scientific publications would be desirable, in principle. The connection to performance rating would present a considerable incentive for the scientists of any discipline to take part in such system. This can lead to a clear improvement of the literature data, which should benefit the respective subject. Therefore, the aim should be, if possible, to collect the literature data for such disciplines in cooperation with an institution that would maintain the data after the rating exercise is completed, and keep them available for scientific users.

The indicators used for the citation analysis in the pilot study for chemistry meet the current international standards of bibliometric research. In the light of the intended effect of research rating, using experimental indicators whose behavior has not yet undergone comprehensive testing is out of the question. However, if it could be shown that novel indicators are more valid and/or manipulation-resistant, the

bibliometric basis should be extended accordingly for future rating exercises. Even in an established rating procedure, there must be some leeway for trialing indicators that are not used in the final assessment. As a special desideratum, indicators should be developed that reflect the reception of compiled volumes and monographs, because limiting the assessment to articles published in periodicals would mean exerting inappropriate influence on the publication culture of the humanities. As was already the case in the pilot study, for future research ratings, too, databases from competing sources would have to be examined for completeness, quality and informative value of the data before awarding a supply contract.

The survey period of five years applied in the pilot study proved to be reasonable. Shorter periods involve the risk of random fluctuations of research performance affecting the ratings, long-term research projects being systematically disadvantaged, and the costs and frequency of ratings becoming excessive.

In the pilot study in chemistry, some institutions found it difficult to make statements about the performance of scientists that had left them before the survey deadline. In sociology, too, where the survey followed the “Current Potential” principle, the retrospective mode of data collection proved to be a problem, at times. In this case, the reason was not that certain scientists had left the institution, but the absence and, under the “Current Potential” perspective, impossibility of institutional procedures for data collection and storage. Consequently, data collectors often had to resort to the notes of individual scientists or working groups. The administrations of several universities stated that, for the future, they would continuously update at least a basic inventory of data in the format used in research rating. Other institutions are already doing so in the context of their regular research reporting. Such data inventory maintained at institution level would mirror the “Work Done At” perspective and reduce the costs of future surveys. Since the data can also be used for self-steering and research reporting of the institutions, as well as for evaluations of any type, they promise additional benefits for the institutions. Therefore, while the steering objectives are better served by a “Current Potential“ assessment, cost considerations would favor a “Work Done At” assessment, which, through the peer review component, can also take into account recent and current changes.

Should the rating procedure be extended to other disciplines, the technical realization of the data collection process should be improved. Making it easier for the subject coordinators to input the required data would raise the acceptance of the entire procedure. The aim should be to develop a modular online system compatible with common file formats, which should also be easily adaptable to the requirements of diverse disciplines.

II.4. Assessment criteria and process

The table of criteria (cf. fig. 2, p. 14), which is a streamlined version of the table originally proposed by the Council, stood the test of the pilot study and should be retained with its essential features.

The criterion “Impact/Effectiveness”, a term that appeared elusive to many users, should be renamed.

Regarding the efficiency assessment, there were suggestions from various parts to include differences in the workload from tasks other than research, e.g. teaching, as well as disparities in the research infrastructures available, and variations in the resources required by different branches within a discipline. If the procedure should be developed further, it should be examined whether these three factors could be taken into account at acceptable cost and, conversely, if statements on the efficiency of research units are actually robust without considering these factors. Possibilities of an empirical weighting of the teaching workload have already been pre-tested and rejected as too expensive. Robust figures for comparison between institutions of the (full) costs associated with the performance of certain research services will not be available for the foreseeable future.

A central desideratum for the assessment of the promotion of young researchers is to obtain data about the subsequent careers of doctoral students and postdoctoral researchers. Improved data of this type would help to arrive at a qualitative assessment of the success of processes to promote young researchers. It would also allow better consideration of successes outside academia. Therefore, universities and non-university institutions should be encouraged to conduct regular alumni surveys.

The efforts of the two assessment boards to agree on a common terminology for the criteria of the “Transfer“ dimension proved to be unnecessary, with hindsight. For the future, the criteria in this dimension should be more closely adjusted to the subject-specific conventions, as the Council already suggested in its recommendations on rankings in the science system.[24]

The five-level rating scale, from “unsatisfactory” to “excellent” was proven, too. The reduction of the scale to three levels for one (in chemistry) or both (in sociology) of the criteria of the transfer dimension made sense for the pilot study, although it affected, in the opinion of some users, the transparency of the results. Consequently, the ratings in the transfer dimension were received with less interest. For the first application of the procedure to a new subject, such reduction might be unavoidable, to begin with. However, indicators of appropriate information value should be developed to allow the same level of differentiation in the ratings of all criteria. In its final report the assessment board for chemistry offered suggestions in this respect. The definition of the criteria in the transfer dimension can also be adapted according to the connections to other societal practices that are typical for the subject and reflected in the availability of indicators. For instance, the assessment board for sociology suggests aggregating the two transfer criteria and argues that this step should also be appropriate for other disciplines of the Humanities.

The pilot study confirmed that peer reviewing is absolutely essential for the reliable assessment of the quality of research performance. This is immediately evident in disciplines like sociology, for which valid citation analyses based on existing data are generally unattainable and where, consequently, the quality assessment requires reading of selected publications without exception. In chemistry, too, the quality assessment by the reviewers essentially depends on qualitative information, such as lists of publications, the self-characterizations of the units, and in many cases the reading of published work, too. Furthermore, the reviewers critically examined the citation indicators and, in some cases, corrected them by resorting to raw data. It comes as no surprise, therefore, that the attempt to produce a statistical prognosis of the assessment results by weighting the quantitative data used in it led to different ratings in 20 – 36 % of all cases: Performance comparisons just based on indicators cannot replace research rating.

---

[24] l.c., p. 47 fn. 46.

The research rating procedure can only be implemented if the assessment workload is shared between the reviewers. The preparation stage with two rapporteurs per rating working independently stood the test of the pilot study. The level of agreement between the reviewers also provided a measure for the reliability of the rating procedure, which turned out high in comparison to the levels known from peer review research.[25] The assessment boards should maintain the option to consult special reviewers, if the board members judge this necessary for the assessment of highly specialized research units or to avoid biases. The rule that the rating proposals of the rapporteurs must undergo plenary moderation and final adoption by the entire assessment board should be kept under all circumstances.

II.5. Results and how they are used

The results of the assessments of both disciplines are highly differentiated, both in terms of making full use of the grade spectrum and regarding the range of ratings any one institution can be awarded according to various criteria. Concerns about possible grade inflation did not materialize. Only 4 – 5 % of all research units were rated as “excellent” for the central criterion, “Research quality”.

The feedback from the institutions and their inquiries about additional data indicate that the institutions are making intensive use of the research ratings. Frequent inquiries about the results of comparable research units at other institutions, in particular, show that the usefulness of the procedure could be further enhanced by publishing the results of every individual research unit. This could not be done in the pilot study, for data protection reasons. Therefore, should the research rating be repeated or extended, personal identification of individual ratings should be precluded by applying appropriate rules to the formation of the research units. If this is not possible for certain subjects or institutions, the consent of participating scientists to the publishing of the rating results should be secured in advance, at the time when the research units are registered. With such precautions, the ratings for each research unit can and should be published after future rating exercises.

---

[25] Cf. A.II.4 and Bornmann & Daniel (2003): “Begutachtung von Fachkollegen in der Wissenschaft“, in Schwarz & Teichler (edts.) “Universität auf dem Prüfstand”, Campus Verlag, Frankfurt: p. 207 – 225; also compare Hartmann & Neidhardt (1990): “Peer Review at the Deutsche Forschungsgemeinschaft“, Scientometrics 19, 419 – 425, where, regarding the funding recommendations, the agreement between reviewers is reported to be at similar levels as seen in the differentiated grade proposals in the pilot study. A high degree of agreement can also be the result of peer pressure: cf. Janis (1982). Groupthink. Psychological studies of policy decisions and fiascoes. Boston: Houghton Mifflin. On measures taken against such effects, cf. A.II.4, p. 19 ff.

Apart from the assessment results, the institutions also received the data reports on which the reviewers had based their assessments. With the percentile values for the quantitative indicators, these reports contain information indicating the relative position of the respective institution in Germany.[26] The inquiries about these data show that the research rating exercise raised the awareness for the importance of continuous data collection at many institutions. One reason for this could be that the survey was conducted in the name of scientific reviewers, not as part of a private ranking initiative.

Apart from the immediate practical benefits of the results for the institutions, the assessed disciplines can also gain productive impulses through a sustained discussion of their subject-specific aims and standards, inspired by quality criteria set in the research rating procedure. In this respect, too, the normative effects of research rating demand continuous, critical monitoring by the subject communities. It would be desirable if the results of this discussion would inspire further development of the research rating system. Considering the learning effects to be expected, it would appear useful to make the basic data collected for the pilot study available for secondary analysis, with data anonymity ensured, especially for evaluation research projects.

Some institutions pointed out that it would be helpful for them – e.g. with regard to staff deployment, third-party funding or publication figures –, if they could directly compare themselves with certain other institutions, in a benchmarking sense. However, this raises concerns that the data could come to be used for a quantitative ranking and their publication would expose the reviewers to pressure to justify their assessments. This could lead to overcautious, conformist assessments and make it more difficult to recruit reviewers. The possible benefits of publishing selected data should be examined, considering these concerns, in dialog with the users.

Although research rating is a retrospective assessment of past achievements, it should also allow statements about the status quo and the development potential of the units assessed. Considering that the data basis already goes back several years, it is important to make this steering information available to the leaders of the institutions as early as possible after the survey deadline. The implementation

[26] cf. fn. 10, p. 15

schedule for future research ratings should be optimized in this regard (cf. II.1, p. 31 ff).

The value of the results would be further increased if repeat assessments could show whether the efforts of individual institutions to improve their research performance were successful. Such effects, however, can arise only some years after the first assessment. Shorter assessment intervals carry the risk of oversteering. The aspect of costs for institutions and reviewers, too, makes an assessment cycle of several years the favorable option. On the other hand, long intervals can lead to a situation where activities are guided by outdated information over excessive periods. International experience, too, favors a cycle time of about five to six years.

II.6. Costs of the procedure

The costs of the pilot study, as a study about a novel procedure, were appropriate. The basic design of the research rating procedure was geared towards transfer to other disciplines, as well as cyclic repetition in each discipline. In this way the costs can be gradually reduced. The margins for cost reductions vary depending on what would be the next steps:

Transfer to other disciplines

Should the rating procedure be transferred to other, previously unrated subjects, the costs per subject had to be expected, initially, to be of the same order as experienced in the pilot study. The expenditure of reviewers' time would be similar, as new indicators would have to be developed for each subject. Related disciplines with similar scientific cultures could profit from each other. Consequently, a flexible system, easier to adapt to other subjects would develop in time, especially once examples from the arts and the technical sciences would be available as well.

As previously unrated disciplines will not be prepared for the data collection stage at the individual institutions, even with the improvements achievable by a longer lead time and possible learning processes in central administrations, the costs for the institutions will not differ significantly, to begin with, from the expenditure required by the pilot study in chemistry and sociology.

Repeat assessment in previously rated disciplines

In contrast, if the procedure was to be repeated in subjects rated before, indicator development would take much less time. Data collection too would become more of a routine process. For such development it is crucial that the scientific institutions collect basic data on their research in central databases in such a way that they can be made available with maximum flexibility. Since the data required for research rating are not unusual, but rather in keeping with international practices for research appraisals, data collection would be made significantly easier by the professionalization of research controlling, which is planned or in progress at many scientific institutions. This process should be further supported by agreements about the survey formats with other institutions collecting data. Also, the costs on the part of the assessed institutions could be reduced considerably if a longer lead time made obsolete or, at least, reduced the need for retrospective surveys.

Both assessment boards have emphasized the importance of careful data cleansing for the quality of the assessment. Therefore, adequate staff capacity for the administrative support of the research rating procedure is essential in order to prevent the reviewers' workload from growing to prohibitive levels. In subjects such as sociology the workload can be reduced drastically by improving the publication data and thereby releasing the reviewers from the task to carry out or mandate checks in many individual cases. The number of research units to be assessed could be significantly reduced in sociology and similar subjects by changing the rules for defining these units in favor of larger organizational units. However, it has to be noted that the validity of the assessment would suffer if the aggregates assessed were anything other than real, viable and relevant units actually engaged in research activities. For the planning of future rating exercises it should be taken into account that the costs for publication and citation analysis were higher than expected.

The procedure for the assessment stage, with the independent pre-assessments by two rapporteurs followed by moderation in the plenum, should be maintained both for an extension of research rating to new subjects and for repeat assessments of the same disciplines. In disciplines where there are no reliable indicators for the reception of scientific publications, direct reviewing of selected publications by reviewers is irreplaceable as a valid procedure for quality assessment. Even if this increases the assessment costs, it is necessary and proper for technical reasons.

Much smoother and more efficient execution of the data collection stage can be achieved by investing in a modular online survey instrument that can be adapted to diverse disciplines. The decision about the development of such system should be made as soon as the research rating procedure is repeated in one of the disciplines rated in the present study.

Cost reductions must not be confused with reductions in the length of the procedure. The latter are not possible without forfeiting the differentiation of the assessment and the allowances for different subject cultures, which are crucial for the acceptance of the procedure. The calls for extended lead times coming from the institutions rather indicate that a less concentrated procedure would help to reduce the overall costs and effort.

When comparing the costs for the research rating system to other evaluation and assessment schemes, one must not forget that no other procedure offers such a differentiated and almost all-embracing comparison of institutions of the public research sector, covering the performance of 9,700 scientists (6,800 FTE) in chemistry and 1,400 (1,160 FTE) in sociology. Considering the scope, quality and degree of differentiation of the assessments produced by research rating and of the ends it is serving, the high costs are generally justified. In the opinion of the steering group, further development of the procedure, with the aim to reduce its costs and, at the same time, optimize the validity and reliability of the ratings is sensible and proper.

Annex: Comparability with published rankings

Among the unique characteristics of the research rating procedure implemented in the pilot study (cf. B.I, p. 27), there are some that forbid direct comparisons of its results with those of published rankings. The main reasons are:

- the differentiation of research rating in several assessment criteria;
- the documentation of internal differences within individual institutions;
- the inclusion of non-university institutions.

Also, discrepancies between the results of different assessment procedures are to be expected because they very rarely refer to the same survey period. Apart from that there are methodical reasons to be taken into account when comparing the results of the research rating exercise with published rankings.

The DFG “Funding Ranking” does not claim to provide a comprehensive assessment of the institutions analyzed for its purposes. However, data like those analyzed for the DFG ranking partly flow into the assessment base for the research rating procedure developed by the German Council for Science and Humanities. When comparing the third-party funding data published by the DFG with those used for the research rating procedure, one must consider the following points:

- Funding Ranking presents a report on approved funds, classifying them by the Review Board and subject to which the respective projects were assigned. Consequently, in some cases figures from the DFG Funding Ranking for chemistry can also include projects of a materials science institute assessed by chemists, but exclude projects of a chemistry institute assessed by physicists. In contrast, the research rating system uses data about the third-party funds spent by certain organizational units that are associated with a discipline through their respective institution.
- In DFG Funding Ranking, the category “SOZ“ embraces social and behavioral sciences, including political science, educational science and psychology. A statement about sociology research at any individual institution is not possible on this basis.

When comparing the current research rankings of the Centrum für Hochschulentwicklung (Centre for Higher Education Development, CHE) for the subjects assessed, considerations must include the following methodical differences:

- The CHE research ranking refers to organizational units that are part of the university courses any subject. For the pilot study, on the other hand, all units engaged in active research in the subject were included. Especially for sociology this resulted in a significantly broader subject area than that of the CHE research ranking.
- Discrepancies in the relative figures, especially the graduation (PhD) rates can be partly explained by actual changes of staff numbers between the survey periods applied by the CHE and the pilot study, respectively.
- The publication figures for sociology cannot be compared directly, because the CHE only publishes weighted numbers of publications. Importantly, the results of the publications searches in the pilot study were checked by the scientists affected. Because of their amendments, the number of documents included in the lists doubled from approx. 5,300 to 10,600. Judging from the experience of the pilot study, significant underreporting of publications is unavoidable without such comprehensive cooperation of the sociologists. This underreporting explains the significantly different results of the publications analysis by the CHE.
- The survey period applied by the CHE is shorter than that of the research rating procedure. This makes the maximum time window for any citation analysis shorter, too. Apart from that, it should be noted that the CHE has not been using subfield-normalized citation data so far.
- To identify so-called “strong research universities”, the CHE feeds absolute and relative publication data, third-party funding and graduation (PhD) data into its calculation. In research rating, these indicators are associated with different criteria. For instance, the CHE does not produce a separate efficiency rating, but includes the relative data in the overall rating.

Comparing the other rankings published in magazines and other periodicals is even more difficult, since the methods of those are usually far from transparent. The ranking of “Focus”, a German newsmagazine, for instance lists a subject-specific “research” value for each university, which, however, is based on data about third-party funds raised, graduation rates and a reputation poll among professors. Without

knowledge of the data it is impossible to explain why, for instance, some institutions that were assessed as having very good impact in the Council research rating were placed in the bottom group by the Focus ranking for 2007.