Part V
The 'Forschungsrating' of the German Council of Science and Humanities. Risks and Opportunities for the Humanities: The Case of the *Anglistik/Amerikanistik* Pilot Study

Rating Research Performance in the Humanities: An Interim Report on an Initiative of the German *Wissenschaftsrat*

Christian Mair

Abstract The author, a professor of English linguistics at Freiburg University, was a member of the German Council of Science and Humanities (*Wissenschaftsrat*) from 2006 to 2012 and, in this capacity, was involved in this advisory body's rating and assessment activities. The present contribution focusses on issues arising in the rating of research output in the humanities and is informed by his dual perspective, as planner and organizer of the ratings undertaken by the *Wissenschaftsrat* and as a rated scholar in his own discipline, English and American Studies.

Over the past decade, rankings—whether home-grown or international—have had a profound impact on higher education in Germany, although the way in which they are being used tends to reveal a degree of tactical short-termism if not downright cynicism. Institutions which come out on top rarely question the procedures by which the welcome result has come about, but are happy to make the most of the free advertising provided. Those not placing so well do not take the result as a motivation for systematic self-study, but rather look to convenient quick fixes which, they hope, will enable them to move ahead in the league tables the next time around.

Within the academic community, rankings have become an informal mechanism of reputation assignment which is not entirely unproblematical but which—at least so far—has had few tangible consequences in terms of structural reform or strategic planning. In wider society, rankings may have some influence on students' and parents' choices of institutions and programmes, though there is as yet no evidence that they are a crucial factor in such decisions, which is probably not a bad thing, either, as the criteria which rankings are based on usually have no very direct bearing on the needs of first-year undergraduates.

In this situation, the German Council for Science and Humanities (*Wissenschaftsrat*), decided to carry out an analysis of the extant rankings in 2004. Its main finding was that the systematic, comparative and often quantitative assessment of research performance had come to stay, but that the methods and criteria employed by the

C. Mair (✉)
English Department, University of Freiburg,
Kollegiengebäude IV, Rempartstr. 15, 79085 Freiburg, Germany
e-mail: christian.mair@anglistik.uni-freiburg.de

M. Ochsner et al. (eds.), *Research Assessment in the Humanities*,
DOI 10.1007/978-3-319-29016-4_15

various rankings were usually not fully transparent and that, moreover, the relevant academic communities had little say in how they were framed (Wissenschaftsrat 2004). The *Wissenschaftsrat's* suggestion for improvement was to develop a rating system in which research output in a particular field would be evaluated comparatively on the basis of criteria developed in consultation with the relevant research community.

As such a rating exercise involved substantial preparation and considerable investment of labour from all parties concerned, pilot studies were deemed essential. The concept was first put to the test on a nationwide scale in the fields of chemistry and sociology—and proved generally workable in both fields, despite their very different objects and methods of investigation (Wissenschaftsrat 2008). Encouraged by this, in 2008 the *Wissenschaftsrat* decided to carry out two further pilot studies, which were supposed to conclude the test phase, and then make the new instrument available on a large scale. The disciplines selected for this second phase of pilot studies were electrical engineering and informatics, on the one hand, and history, on the other. While the engineering pilot was successfully completed in June 2011 (Wissenschaftsrat 2011), the history pilot ended in a deadlock between the *Wissenschaftsrat*, representing the advocates of measuring research output in the humanities, and the Verband deutscher Historiker (Association of German Historians), representing the research community to be rated. As some of the debate was conducted in the culture pages of major national broadsheets, it generated an amount of publicity which, at least for the *Wissenschaftsrat*, was not entirely desirable in such an early phase of testing the new instrument.

On the other hand, it is the high profile that this episode gained which makes it instructive and interesting beyond its immediate academic-political context. In the remarks which follow I shall therefore take it as a starting point for a discussion of the particular difficulties—objective and subjective—surrounding the comparative measurement and evaluation of research output in the humanities and to present the *Wissenschaftsrat's* line of argumentation on this important issue.

In principle, there is no reason why a rating exercise as envisaged by the *Wissenschaftsrat* should be offensive to scholars' sensibilities in the humanities. After all, in its critique of the current situation, the *Wissenschaftsrat* points out the superficiality and lack of transparency of most existing *rankings* and makes the point that any instrument used to measure research performance needs to fit the discipline it is applied to. The *ratings* which the *Wissenschaftsrat* (Wissenschaftsrat 2004, pp. 33–43) suggests as the appropriate alternative are supposed to:

- be conducted by peers who understand the discipline they are evaluating,
- apply criteria specific to the field being evaluated,
- evaluate research output in a multi-dimensional matrix rather than a simple rank list,
- differentiate between achievements of individual 'research units' representing the field at a particular institution.

The last-mentioned criterion in particular should be welcome to scholars in the humanities, who define their research agenda very much as individuals and would resent their achievement to be levelled into departmental averages in a rating exercise.

While the preparation for a rating may involve a certain degree of nuisance and the rewards may be uncertain, the overall design features should find a sympathetic audience among humanities scholars. As a principle, informed peer review is accepted in the humanities as in other academic fields. It determines what gets published or who gets selected for positions, and at conferences or similar forums humanities scholars certainly enjoy the opportunity of showcasing their work and benefit from constructive criticism and advice extended by peers as much as anyone in academia.

What then is the cause of the hostility towards the rating exercise articulated by German historians (or at least their spokespeople in the association)? At least in part, I would contend, the conflict was due to a communication problem. Rankings and ratings, including the *Wissenschaftsrat's*, tend to be presented in a discourse of administrative control and neoliberal new public management which makes many scholars in the humanities suspicious from the very start. Their main experience with this discourse has so far been gained in the defensive rather than the offensive mode. Strategic planning of research has been experienced as increasing regimentation, increasing pressure to produce largely bureaucratic documentation and—in the extreme case—withdrawal of personnel and resources. That the humanities stand to gain from strategic planning—for example through improving career prospects for young scholars or claiming their due place in expensive digital infrastructure projects—has been less obvious by comparison. In this situation, any type of ranking or rating is thus likely to be considered as part of an unhealthy trend towards the bureaucratization, commercialization and commodification of higher education.

Let me briefly illustrate the type of miscommunication I have in mind with one of the *Wissenschaftsrat's* own formulations. Both internally and in several external presentations it has defined the purpose of the rating exercise as 'Unterstützung der Leitungen bei strategischer Steuerung durch vergleichende Informationen über Stärken und Schwächen einer Einrichtung' [supporting administration in its strategic planning by providing comparative information on strengths and weaknesses of a unit] (see Wissenschaftsrat 2004, p. 35, for a published version). Putting things in this way is certainly not wrong, but—in view of what has been said above—clearly not the best way of enlisting the support of the scholars whose participation is required to make the exercise a success. While the formulation allows us to infer the threats that may accrue from under-performance, it is not very explicit on the rewards to be derived from co-operation, both in terms of a particular field and the individual researcher. Researchers in the humanities are generally individualists and therefore sceptical about higher-level strategies of promoting or regimenting their scholarly creativity. They are competitive but not necessarily in the corporate sense of championing their institution. Successful teams are more likely to be composed of scholars working in different places than of colleagues belonging to the same department.

In his public debate with the *Wissenschaftsrat*, Werner Plumpe, the renowned historian and president of the German Historians' Association at the time, emphasizes exactly these points in his critique of the proposed rating (Plumpe 2009). Quan-

tification and standardization, he claims, may suggest the simplicity that political decision makers in university administration and higher-education bureaucracies crave, but this simplicity is a spurious illusion [in his own words (Plumpe 2009, p. 123): 'teilweise quantifizierte, immer aber parametrisierte Informationen für politische Diskussions- und Entscheidungsprozesse, die gemessen an der Realität des Faches unterkomplex [sind]']. An even bigger illusion is the assumption that success in research is the result of stimuli set in the system or advance planning of other kinds ['Illusion, Wissenschaft lasse sich parametrisch durch das Setzen bestimmter Anreize steuern'] (Plumpe 2009, p. 123). According to Plumpe, a standardized rating is not merely useless but counter-productive, because it encourages scholars to focus on meeting the targets of the system rather than the often different standards of professional integrity and scholarly excellence ['Herausbildung und Verfestigung strategischer Verhaltensweisen, die zumindest in den Geisteswissenschaften die akademische Kultur zerstör[en]'] (Plumpe 2009, p. 123). In short, the field of history does not owe it to itself or anyone else to take part in such a problematical project:

> Das Fach habe es aber weder nötig noch sei es im eigenen Interesse verpflichtet, die gefährlichen Illusionen der derzeit politisch hegemonialen Strömungen zu bedienen.
>
> [Neither self-interest nor external necessity forces the community to pander to the current hegemony's dangerous illusions.] (Plumpe 2009, p. 123)

As we see, the opposition is comprehensive and formulated with considerable rhetorical investment. A compromise between the Historians' Association and the *Wissenschaftsrat* was not possible. While the opponents of rating could claim a victory and were in fact heralded as champions of academic freedom in some of the press reportage, the *Wissenschaftsrat* found itself in a bit of a fix. In an atmosphere thus charged, it would have been futile to just move on and approach another field in the humanities to enlist its co-operation. The way out of the impasse was the creation of a working group bringing together a wide range of scholars in the humanities—from philosophy through literature and linguistics all the way to area studies, including the *kleine Fächer*, highly specialized areas of enquiry such as cuneiform studies or Albanology, which in the German system are frequently incorporated as micro-departments consisting of one professor and one or two lecturers or assistants. This interdisciplinary working group was expected to assess the suitability of the *Wissenschaftsrat's* proposed rating to the humanities and suggest modifications where it held them to be necessary.

The present author was privileged to be part of this working group and can testify to the open atmosphere of discussion which made all participants aware of the wide range of research methods and theoretical frameworks found in the contemporary humanities. Most members of the group eventually (though not initially) accepted that rating research output according to the *Wissenschaftsrat's* model was possible in the humanities, might even have beneficial side effects for maintaining and developing quality in the individual fields, and be a means of securing the humanities' general standing in the concert of the other disciplines. Intense disputes, however, arose every time concrete and specific standards of evaluation had to be formulated. Early drafts

of the recommendations contained fairly contorted passages on the relative merits of the traditional scholarly monograph as against the co-authored paper in a peer-reviewed journal, on the need to encourage publication in English while safeguarding the continuing role of national languages as languages of scholarly publication, and so on. About half way through the proceedings, participants realized that the best way to solve these issues for the time being was to defer them, i.e. to state the problem but to expect the solution to emerge from subsequent discussions in the individual research communities concerned. The recommendations thus grew slimmer, but improved from meeting to meeting as discussants realized that they had to aim for a mid-level of abstraction and leave the concrete fleshing out of standards to the discipline-specific experts. In a slight departure from existing *Wissenschaftsrat* rating conventions, the following three dimensions of evaluation were proposed (Wissenschaftsrat 2010, p. 20):

- Forschungsqualität [quality of research]
- Forschungsermöglichung [activities to enable research]
- Transfer von Forschungsleistungen an außerwissenschaftliche Adressaten [transfer of research achievement into non-academic domains].

To accommodate possible slower rates of maturation of research results and slower dissemination and reception, the standard five-year cycle of assessment was extended to seven years. It will be a major challenge to rating exercises based on these recommendations that qualitative measures were prioritized over quantitative ones. Thus, for the assessment of research quality, each 'research unit' will be asked to submit the five publications from a relevant seven-year period which are considered most important. The technical designation 'research unit' is intended to make possible reporting at a contextually appropriate level intermediate between the individual researcher and an institutionalized administrative unit such as a 'department' or an 'institute'. In a traditional German humanities context, this level would typically be understood to be the '*Professur*', i.e. the professorial '*Lehrstuhl*' or chair comprising the professor and his or her assistant(s). Discussions in the working group suggested that some academics would be quite happy to dispense with this intermediate layer in practice and submit five publications per professor, thus defining the relevant unit of documentation as the individual advanced researcher. Clearly, those responsible for the next pilot study will take the opportunity to clarify this contested issue against the background of their discipline.

The most salient feature of the proposed procedure when compared to rating in the natural sciences is that quantitative information, such as number of publications, will play an ancillary role only. This is justified, though, in view of the fact that standard quantitative indicators such as impact factors or citation indices are only marginally relevant in the humanities. One additional dimension of evaluation which it was judged necessary to include in rating research quality similarly defies quantification, namely a researcher's scholarly reputation. In view of reputation's auratic and intangible nature, those members of the working group who would rather not have included it as a criterion will probably take consolation from the fact that it will not have the same importance for all disciplines and certainly not for all individuals.

One of the more convincing ways of measuring reputation was considered to be taking note of the award of prestigious research prizes, such as the German Research Foundation's (DFG) Leibniz Award. Those who advocated considering reputation emphasized that it was not something which lapsed in the seven-year time-window relevant for measuring performance.

The term *Forschungsermöglichung*, not conventionally established, was used as a cover for activities which did not necessarily result in research publications by the principal investigator, but promoted research activities in a wider sense. Typical examples would include contributions to the development and maintenance of important research infrastructures, such as digital text archives or linguistic corpora, acquisition of external funding for research teams providing career opportunities for young researchers, etc. The distinction between the two dimensions of quality and enabling was felt necessary as (a) the mere fact that research in the humanities was funded by external grants did not mean that it was necessarily of high(er) quality and (b) across virtually all humanities disciplines the individual researcher was considered to be in a position to produce first-rate research unaided by teams or expensive infrastructure.

Transfer was expected to take forms appropriate to the individual disciplines, ranging from involvement in exhibitions and museums (art history) via in-service teacher training (foreign languages) to consulting activities (philosophical ethics).

As I briefly hinted at above, it is also very interesting to note the points on which the general recommendations are silent. They do not pronounce on the relative merit of different formats of publication, such as the article in a refereed journal, the article in a volume of conference proceedings, or the monograph. What constitutes an effective or prestigious place of publication is a question for individual disciplines to decide, and linguists' answers will certainly be different from historians'. Personally, I found this attitude of tolerance a little too generous as I am convinced that publishing cultures in all humanities subjects are in a state of transformation. The bad news is that too much is published, and too little is read, but the good news is that in many disciplines informal hierarchies of publishing outlets are emerging which may not be as rigorously enforced as the impact-factor-based reputation hierarchies in the natural sciences, but nevertheless provide orientation to scholars as to where they should strive to publish in order to ensure a maximum audience for their findings.

Another important point the recommendations are silent on is language(s) of publication. Research in the humanities is informed by culture- and language-specific traditions of academic writing, and most scholars in the humanities consider multilingualism an asset in their practice. Arguably, however, our current practices and the academic language policies currently advocated do not promote the most intelligent kind of academic multilingualism in the humanities. Knee-jerk reactions to combat the spread of English and promote academic publication in the respective national languages will usually find favour with the public but are potentially harmful. Consider the following example. A German specialist on the Portuguese language with interesting results on the specificities of Brazilian as against European Portuguese has three theoretical options: (a) publish the findings in German and guarantee dissemination in the peer group most relevant to his or her career, (b) publish in Portuguese

and thus reach the speakers of the language itself, and (c) publish in English to reach the global community of experts on Portuguese. Each of the strategies will potentially lose some readers: people interested in the Portuguese language not reading German (a), general linguists with no particular fluency in Portuguese (b), and people interested in the Portuguese language unable to read English (c). To compound the issue further, the strategy adopted will partly determine the use made of the findings. Publication in German or English will attract additional readers with no specific interest in Brazilian Portuguese as such, but with an interest in the standardization of pluricentric languages in general (e. g. Canadian English vs. United States English, or convergence and divergence between Standard German as used in Austria, Switzerland and Germany). Publication in German may lead to more intensive popularization of the findings among the small group of German-based teachers of Portuguese as a foreign language. These are merely some of the legitimate motivations which guide writers in the choice of languages for publication.

Conceivably, publication in German or Portuguese might also be employed for less than honest purposes, for example as a convenient method to get away with the unreflected use of traditional philological methods by insulating one's work from potential criticism articulated by a now largely English-speaking international community of 'modern' general linguists. But then again, this very Anglophone global linguistic establishment could be accused of cultural imperialism, which for example indeed manifests itself often in refusing to recognize important innovations until they are made available in English. Given the complexity of the politico-linguistic terrain in the humanities, researchers need more support than they are getting now. For example it is much better to fund the translation of excellent work published in languages other than English than to force researchers who are not entirely confident in their language skills to write in English themselves.

The labours of the working group have had one immediate positive result. The group's recommendations have made it possible for the relevant professional associations in the field of English and American Studies to participate in a pilot study. The panel started work in March 2011. Its findings were published in November of the following year (Wissenschaftsrat 2012). The results of the research rating *Anglistik/Amerikanistik* will eventually help determine whether the *Wissenschaftsrat's* approach to measuring research output in consultation with the relevant communities will have a future as a routine tool in the German system of higher education.

If the pilot study turns out to be successful, English and American Studies in Germany will take the rating exercise as the external stimulus to undertake the necessary critical stock-taking that every department needs at intervals. Owing to the safeguards described above, researchers can rest assured that their output is measured against criteria developed by their peers. In the full concert of disciplines in the university, scholars in English and American studies will not have to plead that their subject represents a special case—a strategy which may bring short-term rewards but which is sure to marginalize a field in the long run.

In marketing the rating exercise to the community, both the *Wissenschaftsrat* and the professional associations will be well advised to rephrase the definition

quoted above ('Unterstützung der Leitungen bei strategischer Steuerung durch vergleichende Informationen über Stärken und Schwächen einer Einrichtung') as:

> Unterstützung der Einrichtung bei Standortbestimmung und Weiterentwicklung durch vergleichende Informationen über Stärken und Schwächen der Leistungen der Forscherinnen und Forscher am Ort.
>
> [Supporting the unit in its efforts to assess its position and develop its potential by providing comparative information on strengths and weaknesses of research carried out locally.]

Understood in this way, the rating exercise can become part of a dialogue between scholars and the other stakeholders in the academic system: administrations, funding authorities, other (and sometimes competing) disciplines and, not least, the educated public whose support the humanities need more than other subjects in order to survive and prosper.

If this sounds too good to be true, consider the following three alternative scenarios which might result from a successful pilot study. It is the year 2027, and we are going through the preparations for the second routine rating for English and American Studies in German higher education (after two seven-year cycles: 2014–2020, 2021–2027).

The first scenario is the dystopian one. Status hierarchies and the peculiarly strong German fixation on the professorial chair[1] will still reign supreme, and we will witness a replay of a heated debate which took place in the 2010 meetings of the working group: 'Is my colleague allowed to report a publication by his assistant, just so he can boost his standing in the rating?' Assuming that there are two 'chairs' in English linguistics in a department, the chief motivation of each chairholder to take part in the rating will still be the hope that each one will turn out the better one of the two (rather than both putting on a good show jointly, in the interest of their department and university, and—not least—for current and prospective students). Among the publications reported we will find a 500-page tome titled *Morphologische Kreativität im nigerianischen Englisch: Neologismen aus der Presse*, published in German, by a German academic vanity press, with a subsidy, and a print run of 150, only five of which are sold outside Germany. This notwithstanding, it is cited as a 'magisterial treatment of its topic, well written and with many interesting case studies'.

This, on the other hand, is the utopian scenario. While the pilot rating (2012) stirred up a lot of furore at the time, the first routine exercise in 2020 added modifications to reduce the burden on evaluators and evaluees, thus increasing acceptance in the community. By 2027, ratings have become socially embedded practice in the academic community, including the humanities, and apart from mild irritation caused by the inevitable bureaucratic requirements, the general response is positive – along the lines of 'good thing somebody is taking note of the research we're doing here', 'well, they've politely pointed out the weaknesses that, to be honest, we have been aware of ourselves—in fact, they've given us free expert advice' and 'good thing we know where we stand this time, and good thing we've improved since the last one'.

[1] Consult the web for the collocation 'member(s) of my chair' and observe how much of the material emanates from the .de top-level national domain.

Neither of the extreme scenarios is likely. As an optimist, I hope for a moderately positive reception of ratings in the humanities. Colleagues will actively embrace ratings as an opportunity to showcase their achievement, but, as in the pilot study, researchers will groan at the tedium of compiling the self-report, and this will be echoed by assessors' groans at the tedium of some of the writing they will have to read.

Acknowledgments I would like to thank Dr. Elke Lütkemeier and Dr. Veronika Khlavna, of the *Wissenschaftsrat's* head office, for reading and commenting on an earlier version of this paper. I have profited greatly from their long involvement in the *Wissenschaftsrat's* rating projects.

References

Plumpe, W. (2009). Stellungnahme zum Rating des Wissenschaftsrates aus Sicht des Historikerverbandes. In C. Prinz, & R. Hohls (Eds.), Qualitätsmessung, Evaluation, Forschungsrating. Risiken und Chancen für die Geschichtswissenschaften? (pp. 121–126). Historisches Forum. Berlin: Clio-online. Retrieved from http://edoc.hu-berlin.de/e_histfor/12/.

Wissenschaftsrat. (2004). Empfehlungen zu Rankings im Wissenschaftssystem. Teil 1: Forschung. Köln: Wissenschaftsrat. Retrieved from http://www.wissenschaftsrat.de/download/archiv/6285-04.pdf.

Wissenschaftsrat. (2008). Pilotstudie Forschungsrating. Empfehlungen und Dokumentation. Köln: Wissenschaftsrat. Retrieved from http://www.wissenschaftsrat.de/download/Forschungsrating/Dokumente/FAQ/Pilotstudie_Forschungsrating_2008.pdf.

Wissenschaftsrat. (2010). Empfehlungen zur vergleichenden Forschungsbewertung in den Geisteswissenschaften. Köln: Wissenschaftsrat. Retrieved from http://www.wissenschaftsrat.de/download/archiv/10039-10.pdf.

Wissenschaftsrat. (2011). Forschungsrating Elektrotechnik und Informationstechnik. Einzelergebnisse der Universitäten und ausseruniversitären Forschungseinrichtungen. Köln: Wissenschaftsrat. Retrieved from http://www.wissenschaftsrat.de/download/archiv/1328-11_Ergebnisdarstellungen.pdf.

Wissenschaftsrat. (2012). Ergebnisse des Forschungsratings Anglistik und Amerikanistik. Köln: Wissenschaftsrat. Retrieved from http://www.wissenschaftsrat.de/download/archiv/2756-12.pdf.

‘21 Grams’: Interdisciplinarity and the Assessment of Quality in the Humanities

Klaus Stierstorfer and Peter Schneck

Abstract In their joint contribution, the president of the German Association for English Studies (Deutscher Anglistenverband), Klaus Stierstorfer, and the president of the German Association for American Studies (Deutsche Gesellschaft für Amerikastudien), Peter Schneck, describe the central motivations behind the decision to actively support the pilot study for the research rating of the German Council of Science and Humanities (*Wissenschaftsrat*) despite some fundamental skepticism among the associations’s members. On the basis of five basic propositions—different in each argument—they both insist that the assessment of research quality in the humanities inevitably requires the central involvement of the disciplines assessed in order to reflect on and formulate the central categories, standards and procedures best suited for such assessments. Such a process must take into account the complexity of research processes and results in the humanities whose qualitative dimensions cannot be fully measured by quantitative methods.

1 Rating Research: Who Needs It, and What Is It Good For? (by Klaus Stiersdorfer)

Research rating and ranking is happening now, at least in German academia in my experience, and it has been growing in the anglophone countries, with which I deal professionally, at an alarming pace and as a kind of *menetekel* for whatever other countries may be planning to do in the future. This is why, and here is my *first thesis*, research rating and ranking cannot be avoided at present. If my first thesis is accepted, then it is worth exploring what it looks like at present in the humanities.

K. Stierstorfer (✉)
English Department, Westfälische Wilhelms-Universität Münster,
Johannisstr. 12–20, 48143 Münster, Germany
e-mail: stiersto@wwu.de

P. Schneck
Institute for English and American Studies, Universität Osnabrück,
Neuer Graben 40, 49069 Osnabrück, Germany
e-mail: peter.schneck@uni-osnabrueck.de

M. Ochsner et al. (eds.), *Research Assessment in the Humanities*,
DOI 10.1007/978-3-319-29016-4_16

Most rating and ranking systems I have come across involve any one of the following procedures: peer reviewing of research publications; measuring of quantities of publications; opinion polls on the research reputations of individual institutions and agencies, or any combination of the three. I will not dwell on the latter two as they seem the most obviously inadequate for rating in the humanities, but do want to broach briefly the topic of peer reviewing which is widely seen as the fairest and most reliable tool of the three. The problems I see with it in its current form have, however, to do with fairness and transparency. With most reviewing procedures, the image of the administration of justice attributed to the so-called dark middle ages seems appropriate. There is little transparency in the application of pre-specified criteria; the actual judges (peer-reviewers) are still shielded from the person under review (the defendant) by the inquisitorial screen of anonymity; and the defendant has hardly any means of recourse to plead his or her case when the verdict is negative. This leads to a situation when most researchers in my field, at least where they have the choice, avoid such reviewing processes as the impression (true or not) arising from this black-box juridical system is imputations of favouritism, nepotism and the pursuit of non-scholarly, strategic or political ends under cover of this anonymity. The much-propounded 'blind' or even 'double blind' peer-review really does not mean that justice is iconically blind (as she should be) as to the addressee of her ministrations (projects under review are all too easily attributable in small research communities), but that reviewees are blinded (as they should not be) as to who is their judge and on what grounds their verdict is really passed. Hence, on this ground and many others, my *second thesis* is, current research rating needs improvement if we want to stick to this practice.

How such improvement can be brought about is, of course, the philosopher's stone here, but before its quest is started, the issue of the necessity of rating research in the humanities in the first place must be dealt with. As this is a short statement, the answer suggested here—which is also the prevalent opinion in the Deutscher Anglistenverband and the official position of its presidency and council—is essentially twofold. First, and this is my *thesis number three*, we need research rating because it is there or, more precisely, scholars in the humanities and their societies and associations should get involved in research rating because they are being practiced at the moment; trying to make oneself heard and get involved in establishing the fairest and best practice possible seems reasonable if not logical and unavoidable. Experience has shown that outright refusal to join the discussion does not help to avoid rating and ranking but produces bad, because inexpertly designed procedures.

Why then has research rating been established in the first place? The simple answer is: money. In the progressive commercialization and economization (if that is a word) of our academia, the political focus on money invested in research has been immense, and hence a mechanism for its distribution was sorely needed. On a simple, outcome-oriented economic model, the logical system is to put money where the best outcome is. Hence the idea to measure research outcomes and put most money where the best outcomes can be registered or at least expected. Thus,

research rating is primarily an administrative tool that has to do with investing and distributing limited funds for research. The crux of defining and comparing precisely these outcomes has long been overlooked or neglected. In the most negative reading, the whole process only shifts the problem to another scenario.

Does rating have any benefits for the scholar or researcher in the humanities? My answer is: No, surely not primarily. In a slightly more personal explanation I would stress that I am not interested in knowing whether my colleague X’s new monograph is better than mine, and if so how much on a scale from 1 to 10, neither do I need to know whether colleague Y’s article in a field I am interested in is rated high or low before I read it as the specific questions I bring to it in my specific research context may differ from quality criteria, nor do I have any desire to be informed whether my publications of the last 5 years are to be graded as 5, 6 or 7 on a scale of 1–10. For purposes of orientation which books and articles to look at in the first place, I have sufficient bibliographic and reviewing tools at hand which are well-established and efficient, even if not easily translatable onto scales from 1 to 10. Thus, my *thesis number four* says research rating is next to useless for the purposes of research itself and time spent on it would be immeasurably better spent on such research.

But, if we cannot reasonably avoid research rating at present, and even if it seems pointless for research, can we gather some lateral benefits from it, although it remains primarily superfluous in the eyes of the researcher? Here my *fifth thesis* is yes, research rating could be devised in such ways that a number of collateral benefits might accrue. Again, a lot of creative thinking could and must go into this question, but I only want to focus on one possible aspect here, that is disciplinary self-reflection. By thinking about criteria how quality of research can be measured and understood, scholars in the humanities will be forced to reflect on their current standards and aims of research and how to define them. This process can help individual disciplines to identify where they stand as a discipline and where they might want to be going in the future, as the steering function of rating procedures can hardly be underestimated. While rating may thus be a good thing for initiating and furthering discussions in disciplines and professional associations such as our Anglistenverband, this does not mean that these guidelines agreed on for the entire discipline are really a good yardstick for individual instances of research. Especially in the humanities we know too well that innovative research is, as Thomas Kuhn, Paul Feyerabend and others have argued, all too often not the kind that is immediately recognizable as such by current disciplinary standards.

Conclusion: Although the benefits seem lateral at best, rating of research is nothing that the humanities can easily avoid at the moment, so it seems better to embrace the discussion leading to its implementation with full commitment in the service of the colleagues for whom we speak in our various associations. The search for a fair, transparent and equitable rating system in the humanities may be a quest for the philosopher’s stone, but that does not mean that, under current circumstances, we should not try as best we can.

Thesis 1: Research rating and ranking cannot be avoided at present.
Thesis 2: Research rating and ranking needs improvement if it is to be continued.
Thesis 3: Research rating and ranking is needed because it is there.
Thesis 4: Research rating and ranking is useless for research itself.
Thesis 5: Research rating and ranking can produce collateral benefits.

2 'Weighing the Soul' of the Humanities (by Peter Schneck)

Let me begin with a little historical anecdote: On April 10th 1901, Dr. Duncan MacDougall, a medical researcher from Dorchester, Massachusetts conducted an experiment to determine the physical existence of the soul. Placing six moribund patients on specially designed scales, the doctor tried to quantify the soul by measuring the weight of the patient's bodies shortly before and shortly after their death. Comparing the difference between the two assessments, MacDougall found that each of the patient's bodies lost precisely the same amount of weight, which was around three-fourth of an ounce, or about 21 g. Since he could think of no other explanation for the difference in weight, the doctor concluded that in the moment of death the soul had left the patient's body; thus the soul not only existed, it's weight could also be pinned down rather precisely at 21 g—which is probably less than one would have expected for such a 'weighty' phenomena as the soul given its metaphysical significance throughout our cultural and spiritual history.

While MacDougall's weighing of the soul may be regarded as one of the countless, equally eccentric and futile attempts to measure the immeasurable—an attempt which is symptomatic for a climate of extreme scientific optimism and positivism around the turn of the 19th to the 20th century—it may nevertheless be instructive for understanding the current struggle between those who propose to assess, rate or quantify the quality of research in the humanities with objective methods of weighing and measurement, and those who think that this attempt would amount to a futile 'weighing of the soul'—that is, an absurd, useless and basically misguided exercise.

The anecdote may be instructive in the context of our discussion for more than one reason, but before I turn to the problem of measuring the immeasurable in the main part of my short remarks, let me clarify a few things from the start.

On the one hand, I am talking to you as a humanities scholar whose teaching and research has been subjected to various forms of quality assessment by an extended number of parties: by other scholars, both from my own field and from other neighbouring fields, by various university administrations and committees, by the review boards of various national and international research funding agencies and institutions, as well as by various assessment boards of the federal state and on the national level. Last, but not least, I have also been asked numerous times to assess myself not by mere introspection, but in a more regulated and prescribed form.

Ever since my performance as a scholar became the subject of a standardized questionnaire for the first time in 1984 at a leading American university, quality

assessment in all its different forms has remained an inescapable part of my scholarly and professional existence.

From this perspective of personal experience as an individual scholar, my feelings towards the continuous increase of assessment processes, the growing repertoire of procedures and protocols, as well as in face of the various institutional and public ratings and rankings in which they result—my sentiments in regard to all this excessive monitoring and controlling could best be described by quoting Elvis Costello: ‘I used to be disgusted, now I’m trying to be amused.’

To put it a bit more precisely; even though over the last decades I have come to experience and somewhat grudgingly accept an astounding number of forms of quality assessment and rating processes in the humanities as inescapable, that does not in any way mean I deem them indispensable. On the contrary, as an individual scholar in the humanities, I have increasingly come to doubt and, in fact, severely question both the essential necessity and the positive effect of quantifying ratings and rankings in and for the specific form of research that is being done in the humanities. To put it bluntly: I find it rather hard, if not impossible, to conceive of any process of calculating and expressing in numbers the difference in quality in regard to research in my field that would actually have any impact other than to regulate it (mainstreaming it, prescribing it) by rather artificial measures of comparison.

Thus, the only thing I learned so far from the ongoing and increasing assessment and quantification of research quality in the humanities is this: Whatever can be quantified, will be quantified—and if it hasn’t been quantified yet, it will be quantified eventually. So I agree with my colleague Klaus Stierstorfer that if ratings and rankings are here to stay there is hardly a way to avoid them—but that doesn’t make them more useful or attractive.

As Werner Plumpe, the president of the Association of German Historians has recently argued with considerable gloom, the sheer pressure of and rush towards ratings and rankings may eventually even reach the unquantifiable soul of the humanities: enforcing quantifying methods on central dimensions of research that cannot and should not be measured and expressed by numerical values only.

There are good reasons to accept some of the more convincing arguments that Plumpe brings forth against rating and ranking procedures in the humanities based on quantification, and I easily agree with most of his criticism and scepticism in regard to the uselessness of quantification for the acknowledgement and assessment of research quality in the humanities. There may also be good reason to subscribe to Plumpe’s skepticism that there is a great danger of misinterpretation, or even misuse by third parties, resulting from the suggestive comparability of mere numerical values—something that must be seen as a central concern given the fact that all these numerical values are (increasingly) used as evidence and arguments for the distribution of resources by universities, by the state (both on the federal and the national level) and by third party sponsors like research foundations (both national and international).

And yet there is something slightly uncomfortable and counterintuitive in this well-stated arguments, and even though I share both the reasoning and the sentiment to a certain degree, eventually the conclusions I draw from the current situation are rather different.

In fact, while Plumpe (and the majority of his colleagues in the association of German historians) have emphatically decided not to take part in the preparatory study initiated by the *Deutscher Wissenschaftsrat* (German Science Council), the *Deutscher Anglistenverband* (German Association for English Studies) and the *Deutsche Gesellschaft für Amerikastudien* (German Association for American Studies) have decided to do just that—despite the fact that we share the fundamental scepticism of our colleagues from the history departments about essential aspects of rating and ranking in the humanities per se.

But there are several reasons for this decision, and some of them have already been presented in summarized form by Klaus Stierstorfer. My task in the following parts of these short remarks will be to describe the specific perspective of the association which I represent in respect to the projected study but also in general. This perspective is particularly characterized by the strong interdisciplinary traits of the research that is being done in German American Studies (or more precisely *Amerikaforschung*).

I said there is something counterintuitive or uncomfortable about the complete rejection of the quantification of research quality in the humanities. While there are, as I readily acknowledged, good arguments against quantification as such, these arguments should not (and probably cannot) obscure our perception of the high degree of assessment by quantification that is already in practice in the humanities—in fact, one could argue that it is quantification which dominates the assessment of individual research in the humanities from the very start until the moment when one has successfully become installed by a committee—on the basis of other assessments—as a university professor. In other words, the professional success in the academic field of the humanities is essentially based on ratings and rankings and other accepted assessment procedures within the field. While these procedures are of course not completely based on or expressed in numbers, one cannot overlook or deny the existence and significance of quantification within these assessment practices in the humanities.

This is not meant to be a rhetorical move—I don't think that my colleagues from the history departments would deny the existence of quantification and ranking procedures within their field and as part of their own daily academic practice. Yet while they would readily attest this, they would probably also insist that all this rating and ranking is only done by peers, and based on meticulous and highly reflected methods of reviewing and critical acknowledgment.

However, if there are procedures of assessment involving quantification established in the field as such, it is obvious that the argument against quantification in the humanities is either a universal one—then it either works or it doesn't; and if it does not work because it can never capture the 'soul' that is the real quality of research done in the humanities, then one should drop it altogether: no more grading of research papers, no more graded forms of assessment for doctoral theses on a standard scale (even when using the Latin terms this is still a quantification of quality), no more ranking lists in committees etc.

On the other hand, if the argument is *not* a universal one (and I don't think it is or can be) then the debate should not be about quantification at all, but, rather about consensual standards of comparison and accepted and/or acceptable conditions of

assessment which make the quantified expression of quality not only possible but even desirable for pragmatic reasons (and a number of factors have been named already during our discussions: the sheer increase of scholarship and its ever growing diversity, international competition and funding schemes within the common European research area etc.).

Another aspect that also tends to be neglected in the debate (and I am only talking about the debate about the pros and cons of assessment and quantification of research quality) is the increasing development of new transnational research and study programs, especially on the young researchers level, i.e. joint doctoral programs within the humanities offered and designed by institutions from different countries across Europe. One of the most challenging tasks is to find a common denominator for the assessment and control of the quality of the study programme and the research of the individual researcher. The same is true for international research consortia: there has to be a shared understanding of the quality standards that would guide and make possible the assessment of the research to be conducted. This is an aspect that is of special significance for American Studies as a discipline and a field of research, since in contrast to English Studies (*Anglistik*), American studies has been conceived from the start as a fundamentally interdisciplinary enterprise. In fact, one could argue that American Studies is the name for research done across the boundaries of various disciplines and since its inception this understanding has always led to intense struggles about the proper methodologies, the common concepts, the shared terminology and, last but not least, the commonly accepted standards of quality in research between all participating disciplines.

Therefore, from the perspective of the scientific community involved in research in American Studies in Germany, the participation in the proposed pilot study by the Science Council has both professional, strategic and pragmatic reasons. On the one hand, it presents a calculated step to maintain a central role in the debate and definition of standard criteria and procedures to assess the quality of research done within the discipline. At the same time, it acknowledges the increasing dynamics of collaborative research agendas across disciplines and across national research areas, which are at the heart of the current struggles for standards, criteria and indicators that may be transferable and commonly acceptable at the same time.

In conclusion, one could summarize the motivational aspects that has guided the decision of the DGfA as follows:

- To assure the active participation and indispensable involvement of the field/scientific community in the process of defining standards and criteria of assessment for the quality of research within the field
- To allow for an open and ongoing debate about standards and criteria within the field and across the disciplines ⇒ interdisciplinary research community
- To actively take on responsibility for the development of common standards and criteria
- To make transparent and critically debate existing standards
- To develop common consensual standards across disciplines that meet the requirements and the dynamics of today's interdisciplinary research in the humanities

Let me end with a caveat: The process certainly is not an easy one, and we do not think that we should drop our guard by replacing our healthy scepticism with a naïve trust in the evidence of numbers and graphs. As has been emphasized, the process of arriving at the shared and commonly accepted standards and criteria I talked about can only be a mixture of top-down and bottom-up approaches and perspectives. To return to my initial historical anecdote: Weighing the 'soul' of the humanities should not simply be translated into a question of grams and ounces, nor should the wealth and diversity of humanities research be assessed as a *quantité negligeable*.

Research Rating *Anglistik/Amerikanistik* of the German Council of Science and Humanities

Alfred Hornung, Veronika Khlavna and Barbara Korte

Abstract The pilot study *Forschungsrating Anglistik/Amerikanistik* is the first implementation of the *Forschungsrating* in the humanities. This chapter presents the findings and conclusions of the rating. It consists of three parts: First, the results of the rating, first published in December 2012, are presented, as well as the conclusions drawn by the German Council of Science and Humanities. Second, Alfred Hornung who chaired the review board reflects on the *Forschungsrating* from the point of view of the chair of the review board as well as an *Amerikanistik* scholar. Third, Barbara Korte writes about the *Forschungsrating* from her perspective as a member of the review board and *Anglistik* scholar.

1 Research Rating in English and American Studies (by Veronika Khlavna and Alfred Hornung)

1.1 Introduction

In May 2008, the German Council for Science and Humanities, which provides advice to the German Federal Government and the State (Länder) Governments on the structure and development of higher education and research, decided to extend its pilot studies of research rating in the fields of Chemistry and Sociology to the

A. Hornung (✉)
Department of English and Linguistics, American Studies, Johannes Gutenberg University Mainz, Jakob-Welder-Weg 18 (Philosophicum), Raum 01-597, 55099 Mainz, Germany
e-mail: hornung@uni-mainz.de

V. Khlavna
Research Policy Department, German Council of Science and Humanities, 50968 Cologne, Germany
e-mail: khlavna@wissenschaftsrat.de

B. Korte
University of Freiburg, English Seminar, Rempartstr. 15, 79098 Freiburg, Germany
e-mail: barbara.korte@anglistik.uni-freiburg.de

M. Ochsner et al. (eds.), *Research Assessment in the Humanities*,
DOI 10.1007/978-3-319-29016-4_17

fields of Technical Sciences and the Humanities (Wissenschaftsrat 2008, pp. 11–17). The overall goal was to test the applicability of research rating methods also in the Humanities. The disciplines selected were *Anglistik/Amerikanistik*, which comprises the subfields of English linguistics, English-language literatures and cultures, American Studies, and English didactics.[1] The results of this research rating of *Anglistik/Amerikanistik* were published in December 2012 (Wissenschaftsrat 2013, pp. 271–333).[2]

The pilot study of the research rating in the discipline of English and American Studies builds on the methodologies and criteria of procedure developed in conjunction with the pilot studies in Chemistry, Sociology, and Electrical and Computer Engineering.[3] One of the most important and essential features of the research rating is that its procedure is *explicitly designed by academic standards*. Academic standards for the research rating are guaranteed by male and female evaluators in review boards as well as by the respective academic associations. The responsibility for the first pilot study of the research rating and its further development were in the hands of a steering group consisting of the members of the scientific commission of the Wissenschaftsrat, individual and institutional members of the major science organizations as well as guests from state ministries and the Federal Ministry for Education and Research. As in the previous pilot studies, the steering group entrusted a review board with the implementation of the research rating for English and American Studies. The scientific organizations and professional associations were asked to nominate potential reviewers with an international reputation who could cover the most important subfields. The review board on English and American Studies, chaired by Prof. Dr. Alfred Hornung, consisted of 19 members. The main objectives of the review board were the definition of the field *Anglistik/Amerikanistik* and its subfields, the determination of criteria for application in the review process, the creation of appropriate questionnaires and the eventual assessments.

Based on the assumption that universities and other academic institutions pursue research in their respective fields and beyond, the assessment of research performance in English and American Studies followed the convention established in the other pilot studies and applied multiple criteria of evaluation, each of them specified by several aspects and operationalized by different quantitative and qualitative data.

---

[1] All institutions active in the research of at least one of the defined subfields were able to participate in the research rating of *Anglistik/Amerikanistik*. The time period chosen for the assessment was 7 years (1 January 2004–31 December 2010). To participate institutions had to have existed for at least half of the survey period. No other criteria, such as minimum number of personnel, were determined. As in the previous pilot studies, the response to the research rating was also very high in English and American Studies. 358 participating professors at the reporting date in 2010 represent 94 % of the 379 professors registered by the Federal Statistical Office for Teaching and Research in 'English and American Studies' (see Statistisches Bundesamt 2010, p. 94).

[2] The results of the participating institutions can be found at: http://www.wissenschaftsrat.de/nc/arbeitsbereiche-arbeitsprogramm/forschungsrating/anglistikamerikanistik.html.

[3] See Wissenschaftsrat (2008, 2013).

As in the previous pilot studies, the assessment of the research performance was based on an *informed peer-review process* by expert reviewers. For each evaluated institution, the reviewers received extensive data with quantitative and qualitative information.

In the following, the levels of the research ratings in English and American Studies and the experiences made in the review process will be outlined and explained. Subsequently, the criteria will be described. The last part will give an outlook on further procedures.

1.2 Procedural Steps

As in other disciplines, the implementation of the research rating in English and American Studies can be subdivided into four phases: 1. subject-specific operationalization, 2. collection of data from the institutions, 3. assessment of the data reviewed by the review board, 4. publication of the results and recommendations for the procedure.

1.2.1 Subject-Specific Operationalization

The subject-specific adaptation of the research rating to English and American Studies included the definition of the field and the subfields, the definition of the criteria and the data, the terms for the participation as well as the preparation of the data collection. The definition of the discipline and its subfields in English and American Studies agreed upon by the review board proved to be adequate and manageable. For comparison purposes the established definitions of the subfields (English linguistics, English Studies: Literature and Cultural Studies, American Studies, Didactics of English) should be reused in future research ratings of English and American Studies. At present the adequate assessment of interdisciplinary research is an area of concern. In order to reflect the different roles and profiles of institutions and to identify their strengths and weaknesses, the research achievements in English and American Studies were also evaluated according to multiple criteria (research quality, reputation, facilitating research and transfer to non-university recipients), each of them with differentiating aspects of assessment. These were mostly operationalized by qualitative information. The background information provided by the institutions on human resources and teaching workloads permitted the contextualization of the data with regard to research activities.

1.2.2 Collection of Data from Institution

The collection of publication lists and data in the institutions were based on the *current-potential* principle (the status of performance of actively employed scholars

at a respective institution on the reporting date of 31 December 2010 over the past 7 year period). The *work-done-at* principle was applied in cases where not all relevant data was available at the reporting date (performance of all scholars employed at the given institution in the 7 year period from 01 January 2004 to 31 December 2010). Thus, the data collection was based on the 'hybrid' approach of *current-potential* and *work-done-at*.

The data collection followed three steps: 1. personnel data, 2. publication data and 3. main data collection. In a first step, the institutions classified scholars actively engaged in English and American Studies according to professional positions, and assigned them to the four subfields. Subsequently, the institutions were asked to submit for each professor three exemplary publications from the survey period. In the course of the subsequent main data collection all other data relevant to the assessment were collected.

Except for the exemplary publications, the data of the institutions were collected in online questionnaires.

1.2.3 Assessment of the Data by the Review Board

As in previous pilot studies, the methods and the *informed peer-review* approach proved to be successful. The assessment was carried out in three steps: First, the two reviewers assigned to respective institutions reviewed the publications and data individually and independent of each other for a preliminary assessment prior to the meetings of the review board. At the meetings the review board formed two separate panels to discuss the preliminary results in subfield-specific groups. Thus English Studies: Literature and Cultural Studies joined up with American Studies, English linguistics with Didactics of English. In a final step, all reviews were put to vote in the general meetings of the plenum.

All criteria were evaluated on the level of the subfields to adequately account for the constitution of the field. After a first review of the data and in preparation for the assessment phase, the reviewers of the respective subfields met with the staff from the Office of the German Council of Science and Humanities to develop criteria for a subfield-specific assessment. This procedure allowed an early analysis of the data material and provided an appropriate access for the assessment of the individual subfields. This approach proved to be successful and should be applied in the future with particular attention to the consolidation of the results gained in subfield-specific meetings with the collectively defined criteria in the review board.

The data assembled for the assessment proved to be of different relevance. While the data collected for the assessment of the criteria 'research quality' and 'facilitating research' provided a solid and reliable basis, the assessment of the criteria of 'reputation' and 'transfer to non-university recipients' was less reliable, also due to some incomplete data. In general, the assessment model however worked out and should be retained with respect to the adjustments recommended in the Final Report of the Review Board (Wissenschaftsrat 2013, pp. 219–271). Efficiency measures were not calculated. The background information provided turned out to be helpful for the

qualification and contextualization of the other data. The high degree of agreement between the reviewers in their rating is a strong support for the reliability of the *informed* peer-review process.

1.2.4 Publication of Results

As in the previous pilot studies, the publication of the results consisted of two parts, the result report (Wissenschaftsrat 2013, pp. 271–333) and the institution-based presentation of results. The results are also available online[4] and allow a direct comparison of the institutions on the level of the different criteria for the four defined subfields.

1.3 Criteria

In line with the rating procedure the following criteria were used for the assessment of English and American Studies: 'research quality', 'reputation', 'facilitating research' and 'transfer to non-university recipients'.[5]

1.3.1 Research Quality

Quality of research is of particular importance in the assessment of research performance. Contrary to previous pilot studies, the assessment of the criterion 'quality of research' was primarily based on the assessment of the quality of the publication output. In addition, information on the quantity of the publication output was used. The focus on a qualitative assessment of the publications in English and American Studies was necessary because a citation-based performance assessment of publications does not exist, which is the case in many disciplines of the humanities.[6]

The qualitative assessment of publication performance was primarily based on the reading of the submitted exemplary publications. For this purpose, each professor

[4]The general results are published at www.forschungsrating.de. The results of the participating institutions can be found at: http://www.wissenschaftsrat.de/nc/arbeitsbereiche-arbeitsprogramm/forschungsrating/anglistikamerikanistik.html.

[5]The complete scoring matrix is available at: http://www.wissenschaftsrat.de/download/Forschungsrating/Dokumente/Bewertungsmatrix_ANAM.pdf.

[6]There are many reasons for the absence of citation indexes: lists of books and monographs in publication and citation databases are often incomplete, publications tended to be in German and hence did not figure in international citation databases, collections of essays and anthologies are not systematically evaluated, the number of citations is no clear information on the quality of a publication, since a citation can indicate both an appreciation and a critique of the respective research positions, and finally there does not seem to exist a unanimous opinion on a quality ranking of journals and other publications.

could submit three publications or publication excerpts of max 50 pages. One of the publications could be that of a young academic affiliated with the professorship. This procedure and, in particular, the possibility of considering a publication of young scholars proved to be advantageous. The assessment of publication excerpts, especially those from monographs, proved to be difficult when the reviewers did not know the complete publication. In the future it should be possible to submit the monograph and to mark the section of about 50 pages to be considered in the assessment. The qualitative assessment of the publication lists and their quantitative information (number of publications according to publication types) enhanced the reading of the submitted exemplary publications. The criteria relevant for the assessment of the publications, namely 'importance', 'degree of innovation', 'originality', 'timeliness', 'impact' (national and international), 'quality of research methods' and the range and influence of the research question for one's own discipline as well as for other fields proved to be adequate.

1.3.2 'Reputation'

The assessment of the criterion of 'reputation' was entirely based on qualitative information given for the assessment aspects of 'recognition' and 'professional activities'. The submitted entries for this criterion were very heterogeneous in terms of quality and quantity which rendered its assessment more difficult. The assessment of data given for 'recognition' proved to be especially difficult. Overall, the assessment of 'reputation' as a separate criterion was justified. To improve data quality, the definition of this criterion and its aspects should be more specified in the future, prior to the collection of data.

1.3.3 'Facilitating Research'

The assessment of 'facilitating research' intended to account for activities immanent in academic fields which enable the performance of research in the first place.[7] The evaluation aspects ('third-party funding', 'young talent', 'infrastructure and networks') and data selected for the assessment of this criterion proved appropriate. Particularly the quantitative data and indicators contributed to the simplification and transparency of the ratings.

The data collected for funding sources and the years of the expenditure of third-party funds was relatively unproblematic for the individual subfields. A possibility to optimize the collection of information on third-party funding activities might be the adaptation of the collection principle for the externally funded projects and the expended third-party funds. Since the records covered externally funded projects granted during the survey period on the one hand and the expenditure of third-party

[7] Refer to Wissenschaftsrat recommendations for comparative research assessment in the humanities (Wissenschaftsrat 2013, pp. 345–367).

funds in each year of the survey period on the other, a connection between the two pieces of information was difficult to assess.

The lists of current doctoral dissertations submitted by the institutions proved to be inconclusive. The assessment of these lists was difficult as the successful graduation can actually not be predicted. Accordingly, this data had lesser importance in the assessment process. At the beginning of the review process, the review board had decided not to assess the achievements in the promotion of young talent on the basis of the number of granted PhDs since this figure just provides information about the quantity but not the quality of the young talent. This approach proved appropriate. To allow a more precise assessment of the success of support for the young talent, this information should still be supplemented by quantitative details of completed PhDs in the future. For the assessment of the achievements of the promotion of the young talent, the collected qualitative information (name of the doctoral candidate, name of the supervisor, title and year) of completed dissertations were more important for the assessment process than information on ongoing dissertations.

An adequate assessment of information on networks and research collaborations, in which the reported scholars were significantly involved, was difficult because of the great heterogeneity of the entries and their varied significance. In some cases, major national and international networks, associations and research centres figured next to less significant and informal networks. In the future, this data should be more distinctively described.

1.3.4 'Transfer to Non-university Recipients'

This criterion assessed the contribution of the institutions with respect to research-based knowledge transfer distinguishing between 'personnel transfer' and 'knowledge transfer'. The institutions attributed different meanings to this criterium, so that the quality of the supplied entries varied accordingly. Moreover, the distinction made by the institutions between scholarly activities and those that are more likely attributable to the domain of transfer was not always comprehensible to the reviewers.

Despite the above difficulties and in view of the increasing importance of the transfer of research results, the record and assessment of transfer activities, especially to the non-university recipients, should figure prominently in the future. The distinction of the assessment aspects 'personnel transfer' and 'transfer of knowledge' was not useful since it was not always reflected in the completion of the questionnaire. In future surveys, this criterion should be defined by more distinctive aspects of assessment and more precise survey instructions.

1.3.5 Background Information

Within the scope of the assessment, the background information was used to qualify all other data. The background information provided about institutions and subfields

turned out to be extremely meaningful and helpful. The possibility to describe the local conditions for the evolution of research projects allowed the reviewers to contextualize the specific research activities, in particular the publications. The information on the teaching and examination workload as well as the personnel situation helped to account for the lack of activities in other areas. For an adequate treatment of this information, self-descriptions should be kept and should not exceed a given space.

The information on vacancies in particular was extremely useful. In order to include this information even more systematically in the assessment process as well as to integrate it into the publication of the results, the collection of data needs to be standardized.

Despite the extremely high value of the background information for the qualification of the other data, it proved nevertheless insufficient. In the interest of a more objective consideration of available resources, a separate calculation and assessment of the efficiency should be included in future reviews.

1.4 Conclusion and Outlook

The successfully conducted pilot study of the research rating in English and American Studies shows that an adequate comparative assessment of research performance in the humanities in general, and in English and American Studies in particular, is possible. The research rating is an apt procedure to account for the particular practices of research in the humanities in the context of research assessment. This is reflected in the development and operationalization of the assessment model and in the specification of the survey period. The mode of representation according to subfields and specific criteria offers addressee-oriented information.

In October 2013, the German Council of Science and Humanities proposed recommendations for the future of the research rating (Wissenschaftsrat 2013) and suggested the extension of the research ratings to more disciplines. The experience gained from the research rating in English and American Studies was incorporated into these recommendations. The financing of the implementation is currently under discussion between federal and state governments.

2 Chairing the Research Rating of *Anglistik/Amerikanistik* (by Alfred Hornung)

The research rating *Anglistik/Amerikanistik* (English and American Studies) carried out under the auspices of the *Wissenschaftsrat* formed part of the pilot studies to assess and establish quality standards in the natural sciences and the humanities. Starting out with chemistry and sociology in 2007–2008, electrical engineering and information technology as well as English and American Studies followed in

2011–2012. Recommended by professional associations and based on my record as member of the review board of the German Research Foundation on European and American Literatures I was asked to chair the review board. Acting on the proposals of the Steering Committee of the German Council of Science and Humanities and a subcommittee, which had developed criteria for the assessment of disciplines in the humanities, a group of eventually 19 members from England, Germany and Switzerland was selected from a list of national and international candidates, provided by their professional associations, the German Research Foundation and the Steering Committee of the *Wissenschaftsrat*. The Steering Committee appointed this group of reviewers and entrusted them with the research rating, supported by administrators of the Head Office (Dr. Rainer Lange, Dr. Elke Lütkemeier, Dr. Veronika Khlavna). In the first session the review board decided over the subfields of the discipline of English and American Studies and the procedure and criteria for the evaluation. Eventually four distinct subfields were defined: English linguistics, English literary and cultural studies, American Studies, and English didactics. The separate treatment of English Studies and American Studies as well as the nonrecognition of a subfield of Medieval Studies were the most controversial points in the discussions. The retrenchment of Medieval Studies, which in the past used to be a subject of English linguistics, turned out to be a fact at most universities which had sacrificed both the language and literature of the Middle Ages to new curricula in Bachelor and Masters of English degrees. The argument for the separate evaluation of the American Studies Master advanced by the Americanists was based on the interdisciplinary nature of this field of studies, which in its best representation at the John F. Kennedy Institute in Berlin, comprises the cooperation of literature, linguistics, culture, history, politics, geography and economics of North America. Indeed, the strengths of American Studies in a number of universities are based on the cooperation of these different disciplines, mostly of literature, culture, politics and history. The creation of these four subfields also necessitated an increase of the number of evaluators in American Studies and didactics of English, eventually making for a parity of respectively five colleagues in linguistics, English and American Studies, and four in didactics.

Guided by the previous pilot studies and considering the special features of disciplines in the humanities, the group eventually settled on four main criteria for the evaluation: research quality, reputation, facilitating research, transfer of research to non-university recipients. The report of the *Wissenschaftsrat* specifies the differentiation of aspects and problems in the evaluation of each of these categories. While the assessment of the research quality and facilitating research proved to be reliable categories, reputation and transfer were difficult to assess. This difficulty might also reflect a difference between national standards. North American and British universities are much more interested in communicating their work to their students and the public. Part of this community service is an adequate and comprehensible representation of a discipline and the profile of a department and its personnel. Such promotional activities also serve to attract students in a strongly competitive system of tertiary education. German academics, especially in the humanities, still seem to be hesitant about the promotion of their work and could learn from their English-language

colleagues. An explanation for this hesitancy could also be the often minimal attention and the low status accorded to disciplines in the humanities in the universities as well as in the public perception. The criterion of facilitating research might contribute to a change in this respect. Facilitating research comprises all measures taken to promote the careers of young researchers in the field. Next to the often long and time-consuming processes of directing individual dissertations, the establishment of structured PhD programs for cohorts proved to be very advantageous. This is also reflected in the successful applications for third-party funds, especially in the constitution of research training groups funded by the German Research Foundation or other sponsors. Our review of these very positive achievements also showed that the major research universities profit most from these joint research programs. At the same time the promotion of many PhDs also necessitates the creation of new avenues for jobs outside of academic careers. In this respect, more attention needs to be directed toward transfer activities and to a more pragmatic orientation of doctoral training programs.

This diversification of research and research training also pertains to the self-conception of the four subfields of the discipline *Anglistik/Amerikanistik*. German linguists of the English language have successfully adapted to international standards, which also includes a trend toward publications of articles in journals instead of lengthy monographs. While the monograph still represents the major piece of original scholarship in the humanities and allows scholars also in smaller departments to document their special expertise, the publication of articles gains increasing importance. This move from monographs to articles also reflects the time available for research in most disciplines of the humanities. Next to German Studies, *Anglistik/Amerikanistik* has the highest number of students who pursue academic degrees or want to enter a teaching career in secondary education. Much time is spent in teaching crowded lectures and seminars and grading papers. Many colleagues of the participating universities used the sections of the questionnaire provided for background information, comments about local conditions, to point to the disparity between teaching and research and to the disregard of teaching in the evaluation process.

The coexistence of academic and teacher training curricula also makes for the hybrid nature of the discipline of *Anglistik/Amerikanistik*. On the one hand the subject of 'English' for future teachers unites all four subfields and combines the tasks of linguists, Anglicists, Americanists and didacticians in teaching courses with a focus on teacher training. In most instances only colleagues in the didactics of English do research in this particular area and hence often score highly in transfer to schools and the public. On the other hand each of the four subfields pursues their research interests geared primarily to academic careers and less to teacher training. Historically the common denominator used to exist in the definition of the comprehensively defined discipline of *'Anglistik'* as philology. The study of etymological features of the English language and close readings of great literature basically stressed the competence of the language as a system, and courses as well as research were conducted in German. Starting in the 1980s this situation has changed with an emphasis on the practical knowledge of English and the performance of the language both in the classroom and

in publications. This change was a response to the powerful influence of English and American popular cultures on young people as well as the increasing importance of ethnic minorities, which challenged the mainstream cultures in the English language countries of immigration: Australia, America, Canada and Great Britain, including the former Commonwealth. Consequently the common bond of philology moved into the background and the four subfields further specialized with an emphasis on cultural studies. The formation of new cooperations and exchange programs with international colleagues and institutions intensified these specializations. The call for inter- and transdisciplinary research programs in the universities corresponded with the new application programs of academic sponsors and favoured adequate research activities. Initially the interdisciplinary nature of research and training in American Studies favoured this field, a fact which also figured prominently in the number of successful applications for third-party funds.

An important part of the research rating carried out by the review board under the auspices of the *Wissenschaftsrat* was its acceptance by institutions, colleagues and professional associations. Early on the *Wissenschaftsrat* organized two meetings in Berlin and Mainz for academic and administrative coordinators from each institution to communicate the process of evaluation and assist in the collection of data about personnel, students and research activities. Representatives of the Wissenschaftsrat, Dr. Veronika Khlavna and Dr. Elke Lütkemeier, and I attended the 2011 and 2012 annual conventions of the *Deutscher Anglistenverband* (German Association for English Studies) and the *Deutsche Gesellschaft für Amerikastudien* (German Association for American Studies) as well as the meeting of the *Deutsche Gesellschaft für Fremdsprachenforschung* (German Association for Foreign Language Research) to inform their members about the evaluation process, to gain their support and to listen to their concerns. Apart from questions about the constitution of the review board, the subdivision of the discipline into four subfields or missing ones, such as Postcolonial Studies or Medieval Studies, the strict time-period of 7 years (2004–2010) for the assessment proved to be the most important points. Even the hybrid approach to the evaluation of *current-potential* and *work-done-at* seemed inadequate and colleagues felt that the work of emeriti and the rupture caused by vacancies were not accounted for. Also, the absence of teaching from the criteria of evaluation was criticized. The differences in department structures in terms of personnel and budget, the comprehensive conception of English as one discipline as opposed to separate subfields and their number of representatives were felt to effect the comparative analysis of ratings. A serious concern was the potential usage of the evaluation results by the authorities in the universities and ministries and pursuant repercussions. In spite of these initial reservations, our reports on first results in the 2012 conventions found more acceptable audiences and many of the concerns raised initially proved to be less relevant in the review process. Maybe the knowledge about such evaluations at American universities made for the more ready acceptance of the research rating among the Americanists.

Reservations about the evaluation of a discipline in the humanities were initially also raised by some members in the Steering Committee of the *Wissenschaftsrat*. The presentation of the results, however, reconciled most members with the evaluation

process, especially since it revealed a number of analogies with the previous pilot studies, not least among them the overall average rating in research quality. At the press conference in Berlin in December 2012 journalists addressed results connected with their local universities and the relevance of the results for the discipline and their fields. My work as chair of the review board ended with a report in the general session of the Scientific Commission of the *Wissenschaftsrat* in January 2013. The high number of participants in the *Anglistik/Amerikanistik* research rating, ca. 90 % of all institutions, and the reliable results convinced the members of the Commission that the research rating developed by the *Wissenschaftsrat* could be applied to a discipline in the humanities. The successful completion of the fourth pilot study also led to the installment of, and my participation in a committee charged to prepare the basis for the extension of the research rating to all disciplines in German universities. In October 2013 the *Wissenschaftsrat* discussed the recommendations of this committee and suggested the extension of the evaluation to other disciplines on a regular basis.

The work in the review board over a 2 year period was carried out in a very cooperative and communal spirit and proved to be rewarding. The feedback between the representatives of the four subfields in separate sessions as well as their cooperation in plenary sessions contributed to the speedy conclusion of the research rating and the successful rendition of the report and its communication to our colleagues at the participating institutions. It was a professional pleasure to chair these sessions and share the insights gained from the informed-peer-review of submitted data with reviewers and the participators from the *Wissenschaftsrat*. The basically good national and international status of the discipline *Anglistik/Amerikanistik*, which emerged from the evaluations and which is documented in the report, is a very satisfying compensation for our work. Feedback from the institutions and subfields as well as positive reactions from ministerial and university authorities to the research rating further substantiate its successful application in the humanities.

3 Quo Vadis *Anglistik*? On Rating a Disintegrating Academic Field (by Barbara Korte)

The German Council of Science and Humanities' 2012 review for *Anglistik und Amerikanistik* gave rise to controversial debate in one branch of the field in particular, namely *Anglistik*. This was once the denomination for English Studies, understood as the study of the English language as well as the literatures and cultures expressed in it from the middle ages to the present, as practiced within departments of English. The results of the rating process document how one traditional area in which German scholars used to occupy a leading position has been practically eliminated from English Studies at German universities: Medieval Studies has survived at only a handful of universities, and it seems to be more strongly connected with other disciplines concerned with the period than with English Studies. Conversely, the field of English Studies now comprises many new interests and specializations, and it has

therefore split up in ways that contributed to dissent over the rating process and its categories.

The decision to run the review under the designation *Anglistik* and *Amerikanistik* was discussed in the raters' preliminary sessions and was determined to be the least controversial appellation for the field as a whole. It pays tribute to the fact that *American Studies* has emerged as a strong and highly visible branch within the study of English literatures and cultures, with a distinct profile defined by its region of scholarly interest (the United States, or North America if Canada is included), with specific inter- and transdisciplinary connections, an internationally renowned beacon (the Kennedy Institute in Berlin) and, last but not least, a very active association that promotes the distinct nature of American Studies (although most professorships for American Studies are still situated within departments of English). From the perspective of *Amerikanistik*, a separate rating category was understandably favoured over the alternative, namely to be rated in a joint group with researchers engaged in the study of all other literatures and cultures in the English language, which the assessment lumped together as *Anglistik*: *Literatur- und Kulturwissenschaft* (English literary and cultural studies). It is scholars from the latter group, or *Anglisten* in the narrow sense, who most frequently voiced objections to the separate rating category for *Amerikanistik*. The two other groups in the pilot study, namely English linguistics and English didactics, remained uncontroversial since their profiles are sufficiently distinct from literary and cultural studies in terms of research interests, methodologies and links with other disciplines.

Arguments for the joint rating of *Anglistik* and *Amerikanistik* asserted, firstly, that they still share major interests in and approaches to the study of literature, film and other areas of cultural production, and, secondly, that the separate treatment of American Studies might further promote a profiling of *Amerikanistik* against—and possibly even at the cost of—*Anglistik: Literatur- und Kulturwissenschaft*. This umbrella term also invited critique since it covers a great diversity of interests and subfields that have emerged over the years in non-Americanist English Studies: *Anglistik* (in the narrow sense) has re-invented itself significantly (not without impulses from American Studies), retaining its historical depth (if diminished as regards the Middle Ages) and some of its traditional philological orientations, but significantly expanding and complementing them under the influence of the various 'turns' of the past two or three decades.

The most prominent and consequential changes within *Anglistik* have been effected through the advance (and institutionalization) of Cultural Studies and Postcolonial Studies, for which we have now also established professorships and, in a few instances, institutes. What the *Wissenschaftsrat*'s review understood as 'English' literary and cultural studies was therefore a much bigger and far more heterogeneous bag of scholarship than that of American Studies. It is unsurprising that there were demands to split this bundle up. It was suggested, in particular, that Postcolonial Studies has become so established in the German academic system that it should have been rated on its own, as in the case of American Studies. But how, then, could one name the rest? Could 'British' Studies contain 'Irish' Studies? And where should one stop? Should specializations in Gender Studies also be rated separately? Or

Shakespeare Studies? The research landscape that the rating exercise was expected to chart would have then become too splintered for the results to be significant. In any case, it is undeniable that, if British, Postcolonial Studies *and* American Studies had been treated as one unit, the results for some universities might have been different.

However, the *Wissenschaftsrat's* pilot study did not only point to rifts within literary and cultural studies: The separate rating categories for linguistics and didactics, though less contested, indicate how it is taken for granted that these two areas have drifted apart from literary and cultural studies. Their umbilical connections to English Studies have not been cut, but some of the linguistic research conducted by members of English departments now seems just as closely affiliated with other linguistics or with cognitive studies, while English didactics is strongly connected to that of other foreign languages or with general didactics and pedagogy. Once more, this emphasizes that *Anglistik und Amerikanistik* is a vexed denomination for an academic field that has become increasingly difficult to define because of internal diversification and crossovers with other disciplines. In this respect, the 2012 study with its four groups reflects a state of disintegration that is not of purely academic interest but implies questions of an eminently political nature that affect individual scholars, individual departments and the profile of the entire field. Departments with strong overall ratings will, arguably, have a better standing within their institutions than those with weaker overall results; they might be in greater demand for collaborative projects within their institution, and hence have better chances of acquiring the third-party funding and number of doctoral students that were important criteria in the 2012 pilot study. Within departments, strongly rated subfields might desire to see their symbolic capital matched by a greater share of the budget. Weakly rated professorships might be abandoned in a department in order to strengthen more strongly rated areas, and so on.

Apart from such political consequences, the discipline might also take the rating exercise as an occasion to reflect upon where it is heading: Are we content to see the field of English Studies become increasingly split up? Do we gain or lose by progressive specializations? To what extent can our universities and departments afford or support such specialization? And how should we advise young scholars in terms of career paths? For instance, should and can English Medieval Studies be revived within the German system? It would be unrealistic to assume that the major divisions within English Studies as it currently stands are reversible. American Studies will remain strong, and Postcolonial Studies will not permit itself to be once more reduced to an appendix of 'British' (?) Studies. Yet English Studies as a whole might profit if its internal *connections* became more visible once again. It is not that these connections were not already there: they exist in the form of organizational units (departments of English), in the cooperation of individual scholars, and they are still implemented in courses of study, notably those that focus on English as a school subject. It is no coincidence that, of the rating's four groups, didactics was the only one with a truly integrative approach to 'English' in all its subfields: language, literature and culture, and significantly also across the *Anglistik/Amerikanistik* divide. Current research interests such as Transatlantic Studies, Migration Studies, Transnational

and Globalization Studies also help to bring the branches of English Studies closer together again and to generate new research areas.

The carving up of an academic field into units suitable for rating creates a publicly visible 'image', but it also gives scholars in the field an occasion to reflect upon whether they see themselves—or their subfields—as adequately represented by that image. The image of English Studies created by the 2012 pilot study seems to have aroused more thought about divisions than about the connecting lines and common research interests that prevent the field from falling apart. A reprisal of the exercise should be sensitive to the criticism voiced against the categories used in the 2012 review. And it should introduce criteria that acknowledge not only transdisciplinary research, but also *intra*disciplinary activities and their importance for the future of English Studies.

References

Statistisches Bundesamt. (2010). Bildung und Kultur. In *Personal an Hochschulen* (Fachserie 11 Reihe 4.4). Wiesbaden: Statistisches Bundesamt.

Wissenschaftsrat. (2008). Pilotstudie Forschungsrating. In *Empfehlungen und Dokumentation*. Köln: Wissenschaftsrat. Retrieved from http://www.wissenschaftsrat.de/download/Forschungsrating/Dokumente/FAQ/Pilotstudie_Forschungsrating_2008.pdf.

Wissenschaftsrat. (2013). Pilotstudie zur Weiterentwicklung des Forschungsratings. In *Ergebnisse und Dokumentation*. Köln: Wissenschaftsrat.

Research Assessment in a Philological Discipline: Criteria and Rater Reliability

Ingo Plag

Abstract This article reports on a large-scale peer-review assessment of the research done in English departments at German universities, organized by the German *Wissenschaftsrat*. The main aim of the paper is to take a critical look at the methodology of this research assessment project based on a detailed statistical analysis of the 4,110 ratings provided by the 19 reviewers. The focus lies on the reliability of the ratings and on the nature of the criteria that were used to assess the quality of research. The analysis shows that there is little variation across raters, which is an indication of the general reliability of the results. Most criteria highly correlate with each other. Only the criterion of 'Transfer to non-academic addressees' does not correlate very strongly with other indicators of research quality. The amount of external funding turns out not to be a good indicator of research quality.

1 Introduction

There are some general concerns with regard to attempts to assess the quality of research carried out in public institutions. At the political level, it is, for example, unclear, what the aims of such assessments might be, and who might use them for which kind of decision-making. Furthermore, scholars complain that such assessments involve a great amount of effort, but it is more than doubtful that assessing research leads to higher quality of research. Another big issue is methodological in nature. Different kinds of methodologies are being employed without any clear evidence about their usefulness or reliability.

In spite of these concerns the English departments at German universities decided to participate in a large research assessment organized by the *Wissenschaftsrat*. The assessment was carried out by peers and explicitly aimed at testing the possibilities and problems of assessing research quality in the humanities, and in a philological discipline in particular. The idea that such an assessment might be especially problematic in the philologies arises from the fact that these disciplines are internally extremely

I. Plag (✉)
Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany
e-mail: ingo.plag@uni-duesseldorf.de

M. Ochsner et al. (eds.), *Research Assessment in the Humanities*,
DOI 10.1007/978-3-319-29016-4_18

heterogeneous, with subdisciplines ranging from historical-hermeneutically oriented research to experimental-quantitative approaches, from highly theoretical to thoroughly applied. For this reason, the peers were explicitly asked to critically assess not only the research they had to review, but also the assessment process itself, over the two years of the project.

At the beginning the peers were highly skeptical concerning the assessment criteria and their operationalization. The assessment was supposed to be based chiefly on qualitative instead of quantitative data, and especially the reliability of these qualitative data was called into question.

The aim of the present paper is to address these concerns from an empirical perspective, answering the following research questions:

- How reliable are the judgments made by individual reviewers? How far do different raters agree, especially on criteria that cannot be quantified? Can one trust these ratings?
- What is the relationship between different quality criteria? For example, is it true that the amount of external funding attracted by a researcher is a good indicator of the quality of the research done by this researcher, as is often assumed?

These are empirical questions that can be answered through a quantitative analysis of the judgment data. The group of peers asked the present author to carry out such an analysis and publish the results in pertinent publications. Previous versions of this paper have appeared in German as Plag (2013a, b). The present version also contains some additional analyses.

In the next section I will give some background information about the procedure, which is followed by an analysis of the rater reliability in Sect. 3. Section 4 investigates the relationship between different assessment criteria.

2 Assessing Research Quality in English Departments: Methods and Procedures

This section presents a short summary of the methods and procedures developed and applied in the research rating. A more detailed discussion can be found in the pertinent report by the *Wissenschaftsrat* (Wissenschaftsrat 2012a, b).

As a first step, the peers discussed the division of English studies into pertinent subdisciplines and the categories for the rating. The group agreed to supply ratings according to four subdisciplines or 'sections': English Literature and Culture (ELC), American Studies (AS), Linguistics (LX), and Teaching English as a Foreign Language (EFL). Each section had a similar number of reviewers (19 overall).

With regard to the categories to be rated the peers agreed on four different so-called 'dimensions': *Research Quality*, *Reputation*, *Enablement*, *Transfer*. For each of the four dimensions a number of more detailed criteria were developed. Institutions were then asked to provide certain types of information for each of the criteria.

Table 1 lists the dimensions and the criteria. Table 2 illustrates the kind of information elicited from the institutions (see Wissenschaftsrat (2012a, b) for a complete list and more detailed discussion).

The information provided by the institutions was then rated according to the nine-point scale shown in Table 3.

Each section of each institution was rated by two peers (referred to as 'raters' in the following). Each rater provided their rating independent of the other rater's

Table 1 Rating dimensions and criteria

Dimension	Criterion
Quality	Quality of output
	Quantity of output
Reputation	Recognition
	Professional activities
Enablement	Junior researcher development
	External funding
	Infrastructure and networking
Transfer	Transfer of staff
	Transfer of knowledge

Table 2 Kinds of information

Criterion	Kind of information (selection)
Quality of output	Three self-selected publications per professorship, lists of publications
Quantity of output	Lists of publications
Recognition	Prizes, research fellows
Professional activities	Journal editorship, reviewing, editorial-board-membership
Junior researcher development	Dissertations, habilitations, prizes, job offers
External funding	Projects, money spent
Infrastructure and networking	Networks, research centers, conferences
Transfer of staff	Course offerings, lectures
Transfer of knowledge	Textbooks, other materials

Table 3 Rating scale

Numeric value	Linguistic value
5	Outstanding
5–4	Oustanding/very good
4	Very good
4–3	Very good/good
3	Good
3–2	Good/satisfactory
2	Satisfactory
2–1	Satisfactory/not satisfactory
1	Not satisfactory

rating. The group of peers discussed the ratings in joint meetings of all raters of a pertinent section. Based on this discussion this group decided on the ratings for the four dimensions. The vast majority of these decisions were unanimous. The resulting ratings by the sections were later discussed and approved in a plenary session with all raters from all sections. Occasionally, ratings were revised based on a re-evaluation of some of the arguments that had led to a certain rating. The final report of the group only contained the ratings of the dimensions, not the ratings for the nine criteria.

For the purpose of this paper two data sets were used. The first one (data set A) contains all independent ratings by all raters. This data set allows us to investigate the level of agreement between the two raters and the relationship between the different criteria. The second data set (data set B) contains the ratings for the four dimensions as decided in the plenary session of the group of peers. This data set is used to investigate the four dimensions on the basis of the final ratings.

For the quantitative analysis the above scale was transformed into a 9-point scale with 5 as the highest score and 1 as the lowest with intervals of 0.5. We will use standard statistical procedures, as implemented in the software package R (Core Team 2012).

3 Reliability of the Ratings

3.1 Rater Reliability

The ratings in data set A show a mean of 2.95 (standard deviation: 0.27). An analysis of variance reveals that there are significant differences between raters (*ANOVA*, $F_{(18,348)} = 188$, $p < 0.05$). Such differences are expectable as each rater reviewed a different set of institutions. Figure 1 shows the means by rater (including 95 % confidence intervals), with each rater being represented by a capital letter.

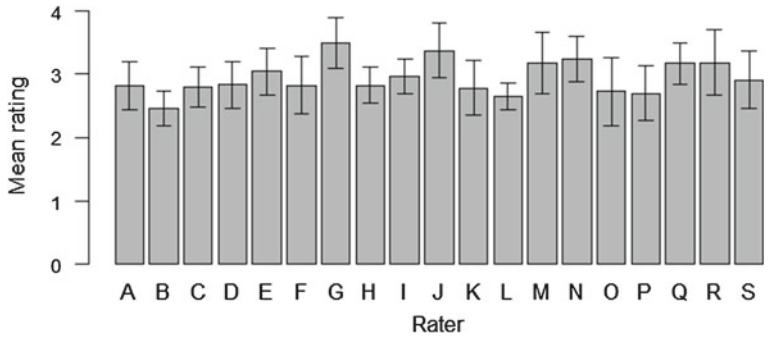
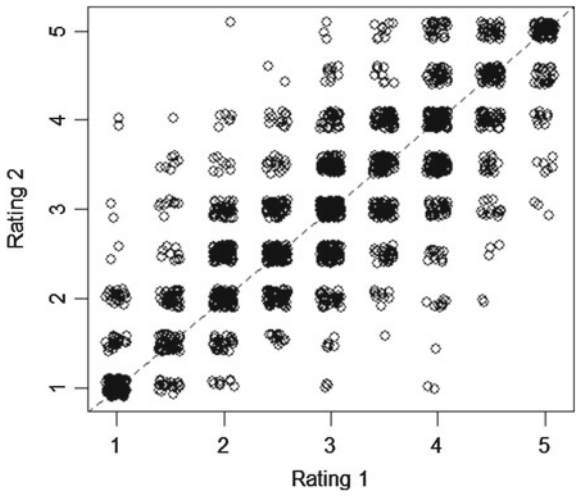


Fig. 1 Mean rating by rater

Let us now turn to the rater pairs and their agreement. 4,110 paired ratings entered our analysis. Figure 2 shows the distribution of the ratings, with some jitter added to each rating for expository purposes. Each of the 2,055 dots in the graph represents one pair of ratings. The scatter is unevenly distributed with most ratings on or close to the diagonal, where the two ratings are identical. Thus we can say that the raters tend to give similar or identical ratings. A look at the differences between ratings corroborates this impression. Figure 3 shows the distribution of the differences between ratings. 40 % of the ratings are identical and another almost 40 % differ only by 0.5. To assess the reliability and consistency of the two raters more formally, we used Cohen's Kappa and Intraclass Correlation (ICC) (see, for example, LeBreton and Senter (2007) for discussion). For our data both measures indicate that there is very strong agreement between two ratings of a given item (Cohen's Kappa: $\kappa = 0.82$, $ICC = 0.802$).

Fig. 2 Ratings by rater



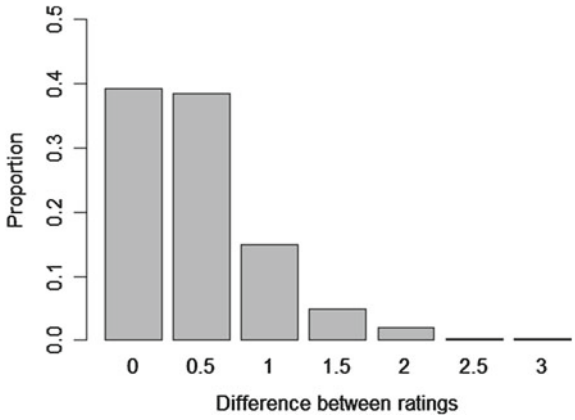


Fig. 3 Distribution of difference between ratings

To summarize, the raters very much agree in their assessment of the criteria, which means that it is obviously possible to reliably assess the quality of research in the disciplines at hand.

It is still an open question, however, whether this reliability differs with regard to the different criteria being rated. This question will be answered in the next subsection.

3.2 Rating Variation Across Different Criteria

An analysis of variance with 'criterion' as independent variable and 'difference in rating' as dependent variable yielded a significant effect of criterion (*ANOVA*, $F_{(12,\ 2012)} = 1.96$, $p < 0.05$). In other words, the difference in the ratings of two raters is dependent on what kind of category was rated. Figure 4 shows the distribution of mean differences by criterion or dimension. Regression analyses show that the six categories with the lowest mean differences do not differ significantly from one another. *Enablement*, however, differs from recognition ($p < 0.05$, $t_{(2012)} = 2.02$) and from all categories to the right of it in Fig. 4.

The dimensions *Research Quality*, *Reputation*, *Enablement*, *Transfer* do not differ significantly from one another concerning the rating differences. With the rating criteria the situation is different. The rating of external funding is least variable, an outcome that is unsurprising given that this criterion is largely dependent on counting sums of money. At the other end of the scale, knowledge transfer seems much harder to reliably evaluate.

It is perhaps striking that the dimension *Research Quality*, which rested primarily on the qualitative assessment of sent-in publications, reached the second best agreement (measured in mean rating difference) in the ratings. This fact can be interpreted in such a way that there are apparently quite clear quality standards in the disciplines

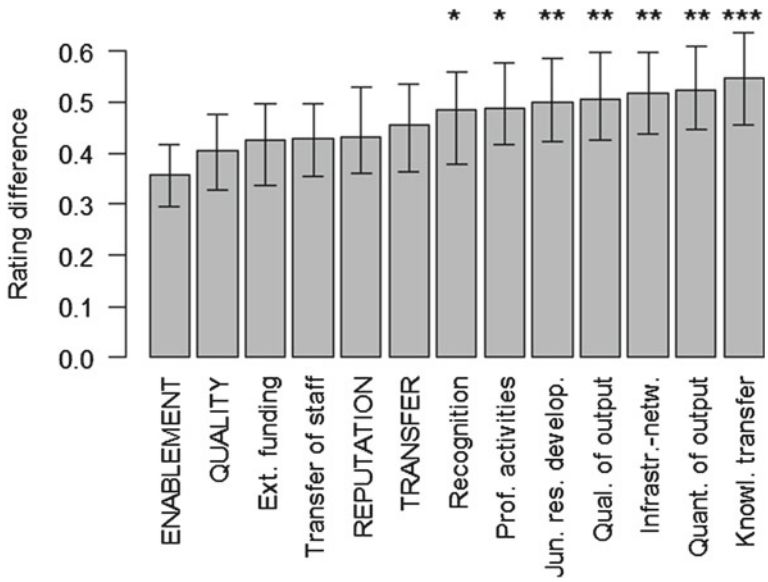


Fig. 4 Mean difference in ratings by category (significance levels for these differences are given by asterisks: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

under discussion, and that these standards were applied by the raters in a consistent fashion.

In sum, there is very good evidence that the peer review procedure as implemented in this project has led to reliable ratings and trustworthy quality assessments.

4 Rating Categories: What Do They Really Tell Us?

In this section we take a closer look at the categories to be rated in order to see in which relation they stand to each other.

4.1 Criteria

If we look at the correlations of the ratings in data set 1 across the nine criteria, we see that all 36 correlations are positive and highly significant (Spearman test). This means that, for a given institution higher scores on one criterion go together with higher scores in any other given criterion. This effect varies, however, quite a bit. Figure 5 illustrates the distribution of the 36 correlation coefficients.

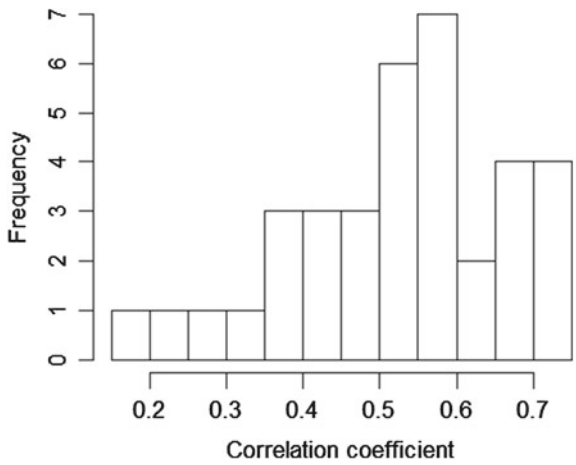


Fig. 5 Distribution of the 36 correlation coefficients for the 9 criteria

Table 4 Highest and lowest correlations between rating criteria

Correlation	Criterion 1	Criterion 2
Strong ($\rho > 0.68$)	Quality of output	Quantity of output
	Professional activities	Recognition
	Professional activities	Infrastructure and networking
	External funding	Infrastructure and networking
	Transfer of staff	Knowledge transfer
Weak ($\rho \leq 0.3$)	Transfer of staff	Quality of output
	Transfer of staff	Quantity of output
	Knowledge transfer	Quality of output

A closer look at these correlations is interesting. Table 4 lists the highest and lowest coefficients.

We can see that some criteria have close relationships to others. A high quality of the publications goes together with a high quantity. This means that people who have very good publications are also the ones that publish a lot. Other very high correlations might be less surprising. That external funds may lead to good infrastructures seems quite predictable, for example.

In the context of today's impoverished universities, external funding has become a prominent issue in political debates inside and outside academia. A common, even if often implicit, assumption in these debates is that attracting external funding is an indication of a researcher's excellence. The present data show that this assumption is not justified. There is a positive correlation between the amount of external funding and the quality and quantity of the research output ($\rho = 0.47$ and $\rho = 0.45$, respectively), but these correlations are not particularly strong. In fact, more than two thirds of the correlations between criteria are stronger.

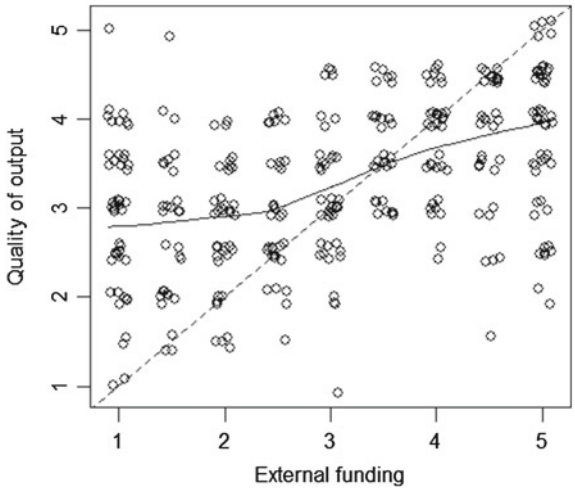


Fig. 6 Quality of output by external funding

Figure 6 shows the relationship between external funding and the quality of the output ($N = 335$, again I have added some jitter). The solid black line gives the trend in the data using a non-parametric scatterplot smoother (Cleveland 1979), the broken line represents a perfect correlation ($\rho = 1$). We can see that the general trend is not particularly strong, at both ends of the x-axis there is a lot of dispersion. What we can say, however, is that high quality research tends to go along with higher amounts of external funding. Conversely, we can state that high amounts of external funding do not necessarily mean high quality research. And there are also two institutions that lack external funding and output top quality research.

These facts suggest that the amount of external funding is not a very reliable way of measuring the quality of research.

4.2 Rating Dimensions

We can apply a similar procedure to data set 2, which contains the final results for the four rating dimensions. Table 5 summarizes the correlation coefficients in a correlations matrix.

All correlations are highly significant ($p < 0.001$, Spearman), but *Transfer* behaves differently from the other three dimensions. Whereas *Research Quality*, *Reputation* and *Enablement* highly correlate with one another ($\rho = 0.73$ or 0.69), *Transfer* does not correlate so well with the other three dimensions (with ρ-values ranging between 0.39 and 0.5). This is also illustrated in the scatterplots in Fig. 7. The left column of panels show the correlations of *Quality*, *Reputation* and *Enablement*, the right column the correlations of *Transfer* with the other three dimensions. The panels on the left show much less dispersion than those on the right, and the trend

Table 5 Highest and lowest correlations between rating criteria

	Quality	Reputation	Enablement
Reputation	0.73		
Enablement	0.69	0.73	
Transfer	0.39	0.49	0.50

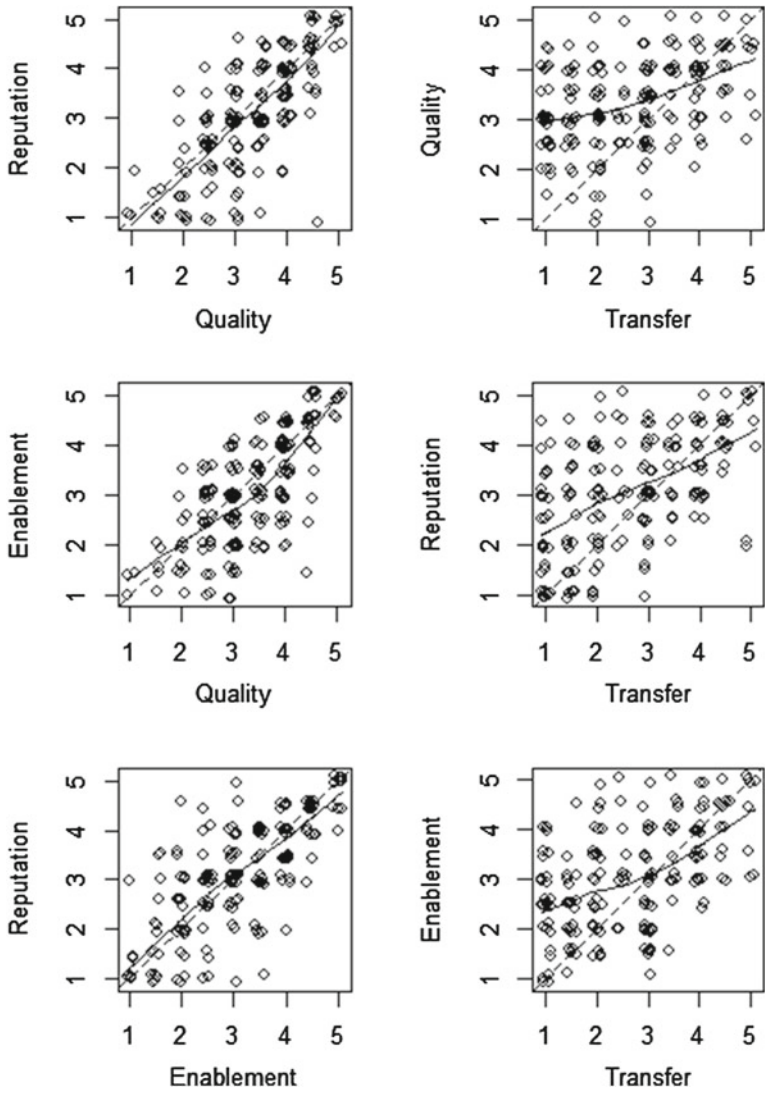


Fig. 7 Relationship between rating dimensions

as shown by the scatterplot smoother in the left panels is also much closer to the diagonal than the one in the right panels.

5 Summary and Discussion

Our analysis revealed that there is strong agreement between raters. This means that the categories to be rated were well operationalized and allowed for a consistent and transparent rating, even if the consistency varied somewhat between categories. It also means that the different subdisciplines represented in English departments in Germany have developed quality standards that are widely shared and that can be used to reach fairly objective assessments of research activities.

With regard to the relationship between the categories three main results emerged. First, there is a significant positive correlation (of varying strength) between all categories. This means that a section of an institution has received similar ratings across the categories to be rated. From a statistical viewpoint this means that the different criteria to a large part reflect the same underlying properties. This was expectable to some extent, but it raises the question of how much effort is actually needed to reach reliable results. The present project involved a considerable investment of time and money, and there is some concern whether such an investment is justified. Politically, the inclusion of many different categories is of course desirable, as it makes the assessment more acceptable for those who are being rated.

Second, not all categories correlate equally strongly, and especially the amount of external funding does not correlate well with measures that directly assess the quality of the research output. This also means that a qualitative evaluation of publications is indispensible for any attempt to assess the quality of research.

Third, we have seen that transfer does not stand in a very strong relationship to other dimensions. This can be interpreted in such a way that transfer to non-academic institutions does not play a prominent role in the research activities of English departments.

Overall we can say that the results of the assessment can be regarded as highly reliable. This result will be to the liking of those that have received good ratings and will be sad news for those who have not reached satisfactory ratings. This brings us to the perhaps decisive question: so what? Or, more concretely, who will use these results and to what end? Who is the addressee of all these assessment efforts?

One might first think of the ratees as primary addressees, as they receive feedback on many aspects of their work. It is highly doubtful, however, whether these scholars need such an assessment in order to learn something about the quality of their research. The scientific community provides constant and ample feedback, either by senior scholars (in the case of dissertations or habilitations, for example) or by peers (in the case of articles, books, jobs, promotion, project funding, prizes etc.), so that all of us seem to get enough feedback to have a fairly good idea about the quality of our own research. Furthermore, for reasons of privacy protection, the present project did not assess research quality at the level of the individual but only at the

level of sections of institutions. The peers were actually sometimes quite unhappy about this restriction since there were sometimes large differences between individuals of one section. These differences then had to be averaged out, which made the assessment less accurate and meaningful than it could have been. For the individual scholar the assessment as done in this project is therefore not really helpful, unless it could be used to improve the situation of an individual section. A reality check of this aspect is sobering, however. While it has happened that universities boasted the achievements of their respective English department as attested in this project on their university websites, I have heard of no tangible increased support (financial or other) accompanying such advertisements.

Let us therefore turn to the other potential addressees of research assessments, i.e. institutions that could use the data for their decision-making (at the departmental, faculty or university level). A discussion of the details of how exactly assessment results may feed into structural or financial decisions taken by university bodies are beyond the scope of this paper, but in general one should be in favour of such decisions being based on trustworthy and reliable data, rather than on the personal biases of decision-makers and their advisors. The present assessment of the research quality of English department certainly provides such a data base.

It should be clear, however, that success in the domain of research is only one criterion for decisions in very complex institutional settings. Apart from information on their research the institutions were also asked to provide information on the institutional settings (e.g. number of students, number of exams, number and structure of staff, number and kinds of study programs etc.). This information clearly indicated that the structural and institutional conditions in many of the departments we assessed are often quite detrimental to the aim of generating excellent research.

References

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*, 829–836. doi:10.1080/01621459.1979.10481038.

LeBreton, J. M., & Senter, J. L. (2007). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852. doi:10.1177/1094428106296642.

Plag, I. (2013a). Forschungsrating der Anglistik/Amerikanistik: Analysen und Reflexionen zur Bewertung von Forschungsleistungen in einer Philologie. *Zeitschrift für Fremdsprachenforschung*, *23*, 177–194.

Plag, I. (2013b). Forschungsrating der Anglistik/Amerikanistik: Analysen und Reflexionen zur Bewertung von Forschungsleistungen in einer Philologie. Anglistik: International Journal. *English Studies*, *24*, 181–194.

R Core Team. (2012). A language and environment for statistical computing. Wien: R Core Team. Retrieved from http://www.R-project.org.

Wissenschaftsrat. (2012a). Ergebnisse des Forschungsratings Anglistik und Amerikanistik. Köln: Wissenschaftsrat. Retrieved from http://www.wissenschaftsrat.de/download/archiv/2756-12.pdf.

Wissenschaftsrat. (2012b). Hintergrundinformation: Pilotstudie Forschungsrating im Fach Anglistik und Amerikanistik. Berlin: Wissenschaftsrat. Retrieved from http://www.wissenschaftsrat.de/download/archiv/hginfo_2612.pdf.