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.

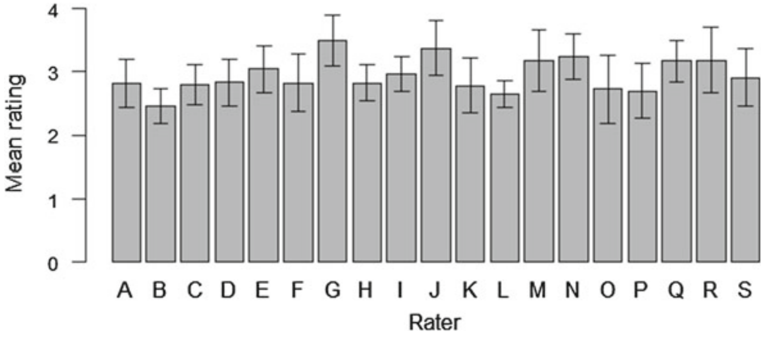
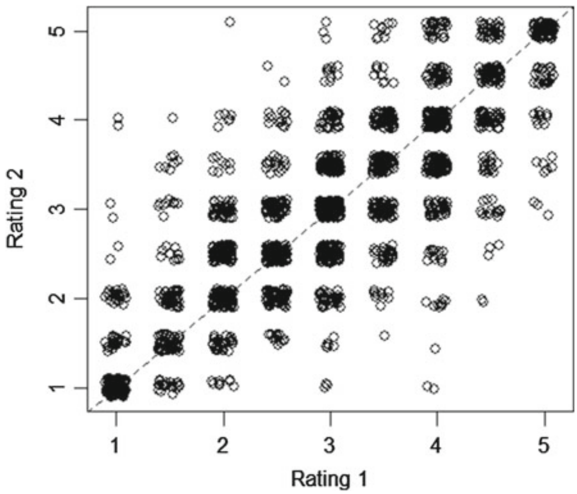


Fig. 1 Mean rating by rater

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



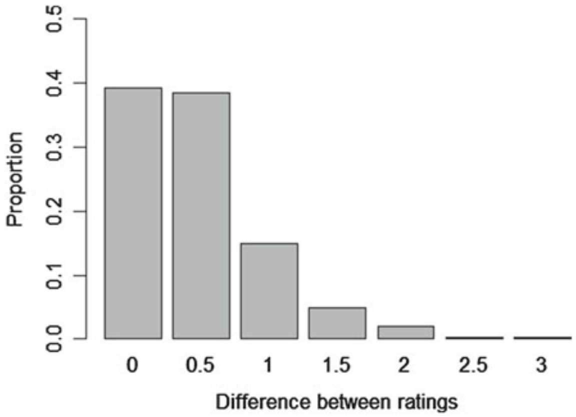


Fig. 3 Distribution of difference between ratings

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

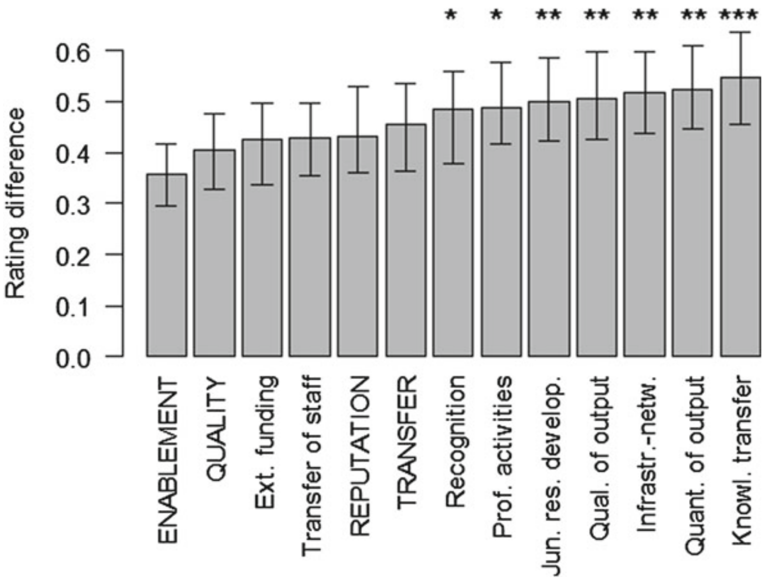


Fig. 4 Mean difference in ratings by category (significance levels for these differences are given by asterisks: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

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.

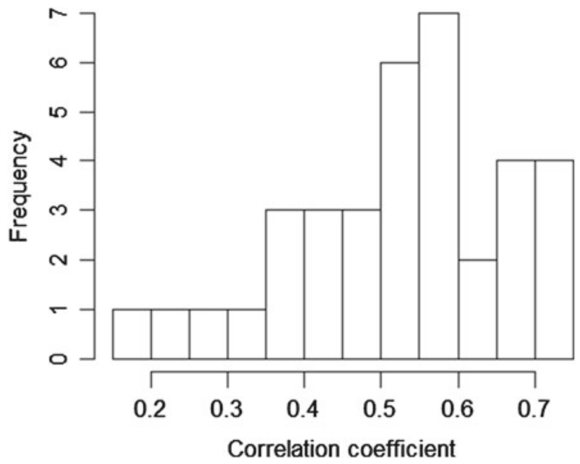


Fig. 5 Distribution of the 36 correlation coefficients for the 9 criteria

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.

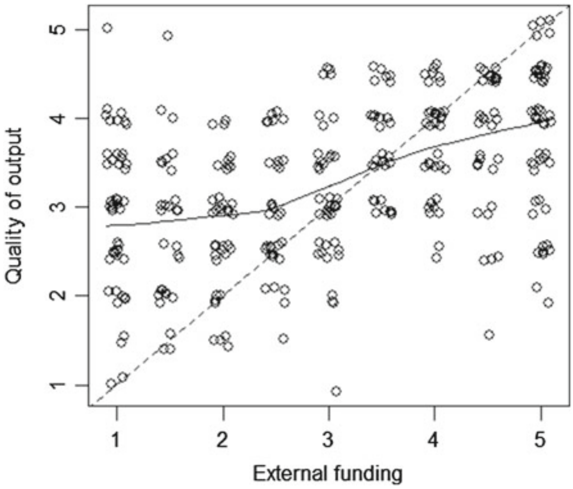


Fig. 6 Quality of output by external funding

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 (ρ = 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

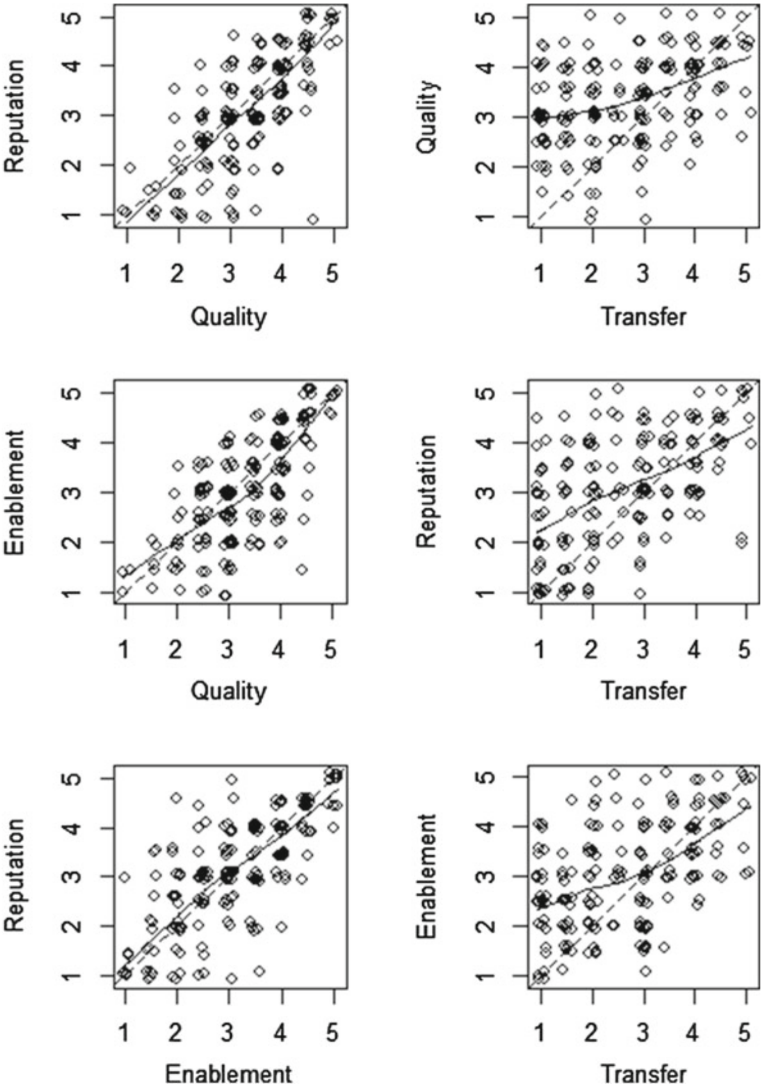


Fig. 7 Relationship between rating dimensions

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.