Forschungsrating der Anglistik/Amerikanistik: Analysen und Reflexionen zur Bewertung von Forschungsleistungen in einer Philologie

Ingo Plag[1]

This article reports on a large-scale peer-review assessment of the research done in English departments at German universities, organized by the *Wissenschaftsrat*. The main aim is to take a critical look at the methodology of the research rating based on a detailed statistical analysis of the 4110 ratings provided by the 19 reviewers. The focus is 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, and 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. Einleitung

Gegen die systematische Bewertung der Forschungsleistungen von Institutionen gibt es vielfältige Vorbehalte und auch in der Anglistik/Amerikanistik ist die Skepsis gegenüber solchen Verfahren weit verbreitet. Auf hochschulpolitischer Ebene ist beispielsweise unklar, zu welchem Zweck die Daten erhoben werden und wer sie dann für welche Art von Entscheidung nutzen wird. Es wird auch angeführt, dass derartige Unternehmungen durch die dadurch gebundenen Ressourcen Forschung eher behindern denn befördern. Außerdem fehle der Nachweis, dass durch Forschungsratings sich die Qualität von Forschung steigern lässt, was aus der Perspektive der Wissenschaft das vorrangige Ziel sein müsste. Berechtigte Fragen werden auch in Bezug auf die vielfältigen methodischen Probleme vorgebracht, die solche Erhebungen mit sich bringen.

Die Anglistik/Amerikanistik in Deutschland hat sich trotz durchaus vorhandener Bedenken zur Teilnahme an einem vom Wissenschaftsrat durchge-

1 Korrespondenzadresse: Prof. Dr. Ingo Plag, Heinrich-Heine-Universität Düsseldorf, Institut für Anglistik und Amerikanistik, D-40204 Düsseldorf, Tel. 0211 81-12963, ingo.plag@uni-duesseldorf.de

führten, auf peer review beruhenden Rating ihrer Forschungsleistung entschlossen. Das Verfahren hatte auch als ein erklärtes Ziel, am Beispiel der Anglistik/Amerikanistik zu testen, wie und ob die Forschungsleistungen einer Geisteswissenschaft sinnvoll bewertet werden können. Daher war auch die Gutachtergruppe von Anfang an aufgefordert, das Verfahren kritisch zu begleiten und sie ist dieser Aufforderung in jeder Phase des insgesamt knapp zwei Jahre dauernden Verfahrens nachgekommen. Insbesondere im Hinblick auf die Methoden bestanden in der Gutachtergruppe anfänglich große Zweifel, ob und wie sich in Anschlag zu bringende Kategorien und Bewertungen operationalisieren und konsistent anwenden lassen. Insbesondere die Verlässlichkeit und Nachvollziehbarkeit von Bewertungen erscheint bei qualitativen Daten, wie sie in der Erhebung im Vordergrund stehen sollten, nicht selbstverständlich. Auch die Natur und Aussagekraft der ausgewählten Kategorien zur Messung der Forschungsleistungen selbst kann mit einigem Recht kritisch gesehen werden.

Ziel dieses Beitrags ist es, einige der kritischen Fragen, die sich aus den genannten Überlegungen ergeben, aus der Innenperspektive der Gutachtergruppe zu beantworten und einer interessierten Öffentlichkeit vorzustellen. Die folgenden Fragen sollen dabei im Mittelpunkt stehen:

- Wie verlässlich sind die Bewertungen einzelner Bewerter? Decken sich die Eindrücke und Bewertungen des einen Bewerters mit denen eines anderen Bewerters, insbesondere bei den nicht quantifizierbaren Bewertungsaspekten? Kann man den am Ende erzielten Urteilen vertrauen?
- Welche Zusammenhänge gibt es zwischen verschiedenen Bewertungskategorien? Sind z.B. Drittmittel, wie vielfach von interessierter Seite angenommen, ein guter Indikator für Forschungsqualität?

Diese Fragen sind, jenseits aller politischen Einstellungen, genuin empirische Fragen, d.h. Fragen, die sich durch eine genaue Analyse der vorliegenden Daten beantworten lassen. So lässt sich beispielsweise durch statistische Verfahren messen, inwieweit Gutachter in ihren Einschätzungen übereinstimmen oder eben nicht. Die Gutachtergruppe beauftragte den Verfasser, der ebenfalls Teil der Gutachtergruppe war, die dafür notwendigen statistischen Analysen durchzuführen. Die Ergebnisse dieser Analysen sind in summarischer Form auch in den Abschlussbericht der Gutachtergruppe eingegangen (Wissenschaftsrat 2012a). Die Gutachtergruppe war allerdings der Meinung, dass eine detailliertere Darstellung in den Fachorganen der

Anglistik/Amerikanistik wünschenswert und notwendig sei, um eine informierte und sachgerechte Diskussion in der Fachöffentlichkeit wie in der allgemeinen Öffentlichkeit zu befördern. Daher wurde dieser Beitrag zeitgleich an den Anglistenverband, an den Amerikanistenverband und an die Deutsche Gesellschaft für Fremdsprachenforschung zur Veröffentlichung in den jeweiligen Organen ihrer Verbände gesandt.

Im folgenden Abschnitt wird zunächst ein Abriss über den Ablauf der Bewertung und die verwendeten Bewertungskategorien gegeben. Abschnitt 3 untersucht dann die Verlässlichkeit der Bewertungen und Abschnitt 4 präsentiert die Zusammenhänge zwischen den Bewertungskategorien.

2. Forschungsrating der Anglistik/Amerikanistik: Methoden und Ablauf

Der Ablauf des Ratings und die Bewertungskategorien werden im Abschlussbericht sowie im Ergebnisbericht der Gutachtergruppe detailliert beschrieben und kritisch diskutiert. An dieser Stelle soll daher nur ein kurzer Abriss gegeben werden, wie er zum Verständnis des Ratings als Ganzes und zum Verständnis der statistischen Analysen im Besonderen notwendig ist. Bei einer solchen Darstellung können notwendigerweise viele Diskussionspunkte nicht adäquat dargestellt werden. Der interessierte Leser wird daher gebeten, für eine ausführlichere Diskussion die Berichte des Wissenschaftsrates zu konsultieren (Wissenschaftsrat 2012a, 2012b). Die Gutachtergruppe verständigte sich zunächst über die zu bewertenden Teilbereiche und die Bewertungskategorien. Nach ausführlicher Diskussion einigte man sich darauf, eine Einteilung des Fachs in die Teilbereiche Anglistische Literatur- und Kulturwissenschaft (im folgenden ALK), Amerikastudien (AME), Englische Sprachwissenschaft (ESW) und Fachdidaktik English (FDE) vorzunehmen. Die Gutachtergruppe war so zusammengesetzt, dass für jeden dieser Teilbereiche eine annähernd gleiche Zahl an Gutachtern zur Verfügung stand (insgesamt 19).

Bei den Bewertungskategorien dienten die drei früheren Pilotstudien des Wissenschaftsrates (Chemie, Elektrotechnik und Soziologie) sowie die Empfehlungen zur vergleichenden Forschungsbewertung in den Geisteswissenschaften des Wissenschatfsrats (Wissenschaftsrat 2010) als Orientierung. Am Ende einer langen, gewissenhaften Abwägung der vielen möglichen Kategorien einigte sich die Gutachtergruppe auf vier sogenannte ‚Bewertungskriterien': Forschungsqualität, Reputation, Forschungsermöglichung

und Transfer an außeruniversitäre Adressaten (im Folgenden kurz Transfer). Die vier Kriterien wurden für die Erhebung aufgesplittet in sogenannte 'Bewertungsaspekte', und für jeden Bewertungsaspekt wurde festgelegt, welche Arten von Information von den Einrichtungen der AA für eine spätere Bewertung eingeholt werden sollten.

Tabelle 1 listet die Bewertungsaspekte und Bewertungskriterien, Tabelle 2 einige wichtige Arten von Daten, die für die jeweiligen Bewertungsaspekte erhoben wurden (eine vollständige Liste der Datenarten sowie deren Diskussion s. Wissenschaftsrat 2012a, 2012b).

Bewertungsaspekt	**Bewertungskriterium**
Qualität des Outputs	FORSCHUNGSQUALITÄT
Quantität des Outputs	
Anerkennung	REPUTATION
Professional Activities	
Nachwuchsförderung	FORSCHUNGS-ERMÖGLICHUNG
Drittmittel	
Infrastrukturen und Netzwerke	
Personaltransfer	TRANSFER
Wissensvermittlung	

Tabelle 1: Bewertungsaspekte und Bewertungskriterien

Bewertungsaspekt	**Art der Information (Auswahl)**
Qualität des Outputs	Drei selbst ausgewählte Publikationen bzw. Publikationsausschnitte pro Professur, Publikationslisten
Quantität des Outputs	Publikationslisten
Anerkennung	Preise, Gastwissenschaftler
Professional Activities	Herausgeberschaften von Zeitschriften, Gutachtertätigkeit, Editorial-Board-Mitgliedschaften
Nachwuchsförderung	Promotionen, Habilitationen, Auszeichnungen und Preise (u.a. Rufe)
Drittmittel	Bewilligte Drittmittelprojekte, Verausgabte Mittel
Infrastrukturen und Netzwerke	Netzwerke, Verbünde, Zentren
Personaltransfer	Weiterbildungsangebote
Wissensvermittlung	Lehrbücher, andere Materialien

Tabelle 2: Art der Information

Die von den Einrichtungen gelieferten Informationen zu den Bewertungsaspekten wurden gemäß einer 9-stufigen Skala bewertet (vgl. Tab. 3).

Skalenwert	**sprachliche Bewertung**
5	herausragend
5-4	herausragend/sehr gut
4	sehr gut
4-3	sehr gut/gut
3	gut
3-2	gut/befriedigend
2	befriedigend
2-1	befriedigend/nicht befriedigend
1	nicht befriedigend

Tabelle 3: Bewertungsskala

Jeder Teilbereich einer Einrichtung wurde von zwei Gutachtern ('Bewertern') unabhängig voneinander bewertet. Diese Bewertungen wurden in gemeinsamen Sitzungen der Gutachter eines Teilbereichs diskutiert und jede Bewertung für die Bewertungskriterien (Forschungsqualität, Reputation, Forschungsermöglichung und Transfer) von der Mehrheit der Anwesenden (meist einstimmig) beschlossen. Diese in den Teilbereichssitzungen zustande gekommenen Bewertungen wurden später im Plenum noch einmal überprüft und abschließend von der gesamten Gutachtergruppe bestätigt oder entsprechend verändert. Nur die vier übergeordneten Bewertungskriterien (und nicht alle neun Bewertungsaspekte) wurden in die publizierte Darstellung der Ergebnisse des Ratings übernommen.

Für die in diesem Aufsatz vorgelegte Studie wurden verschiedene Datensätze untersucht. Ein Datensatz ('Datensatz A') umfasst die unabhängig voneinander abgegebenen Bewertungen der Bewertungsaspekte durch einzelne Bewerter. Dieser Datensatz ermöglicht eine Untersuchung der Übereinstimmung der Bewerter sowie der Zusammenhänge zwischen einzelnen Bewertungsaspekten. Ein weiterer Datensatz ('Datensatz B') enthält die Bewertungen für die übergeordneten Bewertungskriterien, wie sie in den abschließenden Beratungen für alle Einrichtungen und Teilbereiche im Plenum der Gutachtergruppe beschlossen wurden. Dieser Datensatz erlaubt eine detaillierte Analyse der Bewertungskriterien auf der Basis der abschließenden Bewertungen.

Für die Auswertung wurden die Bewertungen in eine neunstufige Intervallskala umgewandelt, mit '5' als höchsten Wert und '1' als niedrigstem

Wert und Intervallschritten von 0,5. Bei der Auswertung kamen standardmäßige statistische Verfahren der deskriptiven und Inferenzstatistik zum Einsatz. Verwendet wurde das frei erhältliche Statistikpaket R (R Core Team 2012). Den üblichen wissenschaftlichen Standards folgend (und für Leser, die in der Statistik zu Hause sind), werden die statistischen Befunde entsprechend dokumentiert. Statistisch weniger versierte Leser können diese technischen Einzelheiten getrost ignorieren und sich auf die nicht-technische Darstellung der Ergebnisse konzentrieren.

3. Verlässlichkeit der Bewertungen

3.1 Übereinstimmung zwischen den Bewertern

Wenden wir uns zunächst den einzelnen Ratern und deren Übereinstimmung zu. Die Mittelwerte der einzelnen Rater schwanken um den Wert 2,95 (mit einer Standardabweichung von 0,27). Eine Varianzanalyse ergab, dass es zwischen den Ratern signifikante Unterschiede in den Bewertungen gab ($p < 0.05$, $F = 1.96$, anova). ‚Signifikant' heißt hier (und anderswo in diesem Beitrag), dass die Verteilung der Daten nicht auf Zufall beruht.[2] Dieses Ergebnis ist auch erwartbar, da jeder Rater eine andere Menge von Einrichtungen zu begutachten hatte. Abbildung 1 zeigt die Mittelwerte der Rater, wobei die Namen der Rater aus Datenschutzgründen durch beliebige Buchstaben ersetzt wurden. Die Säulendiagramme sind durch Konfidenzintervalle ergänzt. Stark überlappende Konfidenzintervalle sind ein Hinweis darauf, dass der Unterschied zwischen zwei Messwerten nicht auf Zufall beruht (vgl. z.B. Gries 2008: 129f).

2 Die dokumentierte Wahrscheinlichkeit *p* gibt an, wie hoch die Wahrscheinlichkeit ist, dass das Ergebnis auf Zufall beruht.

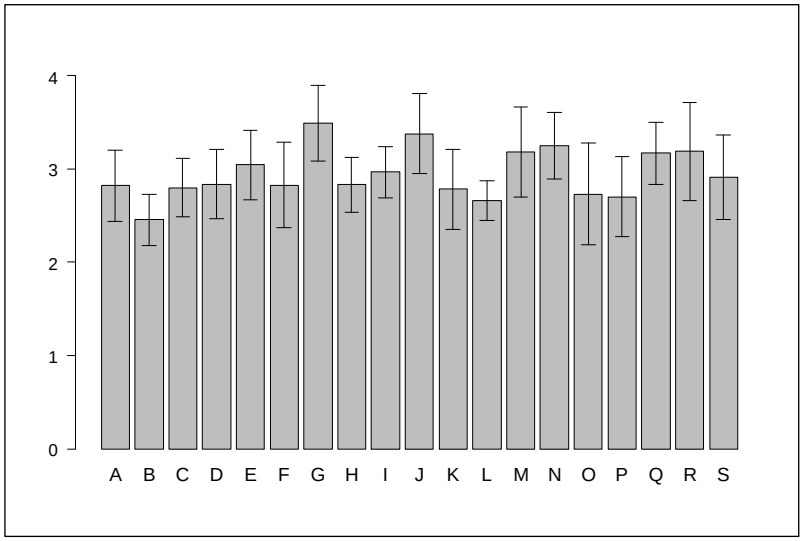

Abbildung 1: Mittelwerte der Rater

Um nun die Übereinstimmung zwischen den beiden Ratern eines bestimmten Teilbereichs einer bestimmten Einrichtung zu untersuchen, wurden die zum Zeitpunkt des Datenabzugs vorliegenden 4110 abgegebenen Ratings einem Korrelationstest unterzogen. Die beiden Bewertungen wiesen eine hoch signifikante positive Korrelation auf (rho = 0,8, p < 2,2e-16, Spearman). Mit anderen Worten, wir finden eine sehr hohe Übereinstimmung zwischen den beiden jeweils abgegebenen Ratings. Abbildung 2 zeigt die Streuung der Bewertungen. Jeder Punkt stellt eines der 2055 Ratingpaare dar, wobei zur besseren Sichtbarmachung der jeweiligen Häufigkeiten die Punkte um den eigentlichen Wert herum leicht gestreut sind. Die größte Häufung von Punkten sieht man auf der Diagonalen, die Bewertungen repräsentiert, wo beide Rater unabhängig voneinander die gleiche Bewertung abgegeben haben.

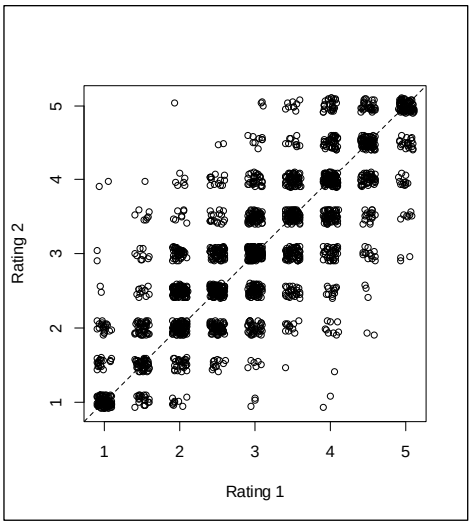

Abbildung 2: Beziehung zwischen den beiden jeweils abgegebenen Ratings

Eine Analyse der jeweiligen Differenz zwischen den beiden Ratings zeigt, dass etwa 40 Prozent der Ratings völlig übereinstimmen und weitere fast 40 Prozent nur um einen halben Punkt differieren. Abbildung 3 zeigt die Verteilung der Unterschiede der beiden Ratings.

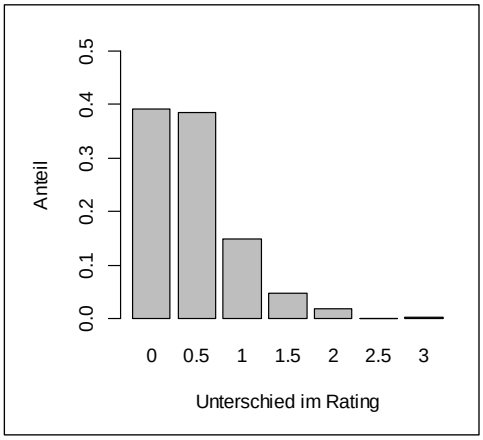

Abbildung 3: Verteilung der Unterschiede im Rating

Insgesamt ergibt sich also ein relativ klares Bild. Die beiden Bewerter stimmen in ihren unabhängig voneinander abgegebenen Urteilen sehr stark überein, es besteht offenbar große Einigkeit darüber, wie die vorliegenden Informationen jeweils zu beurteilen sind.

Es stellt sich aber die Frage, ob diese Übereinstimmung bei allen Kategorien gleich groß war, oder ob es Kategorien gab, die nicht so eindeutig zu bewerten waren wie andere. Dieser Frage geht der nächste Abschnitt nach.

3.2. Bewertbarkeit der Informationen nach Kategorien

Eine Varianzanalyse mit 'Kategorie' als unabhängiger Variable und dem Unterschied zwischen den Ratings als abhängige Variable ergibt insgesamt einen signifikanten Effekt für 'Kategorie' ($p=0,02$, $F(12)=1,96$, anova). Mit anderen Worten, der Ratingunterschied ist abhängig von der jeweiligen Kategorie, die zu bewerten war. Abb. 4 zeigt die sortierte Verteilung der mittleren Unterschiede zwischen den beiden Bewertern nach Kategorien. Eine Regressionsanalyse mit den genannten Variablen zeigt, dass sich die sechs Kategorien mit den niedrigsten Unterschieden nicht signifikant voneinander unterscheiden, wohl aber 'Forschungsermöglichung' von 'Anerkennung' ($p < 0,05$, $t(2012)=2,02$) und von allen weiteren Kategorien rechts davon in Abb. 4.

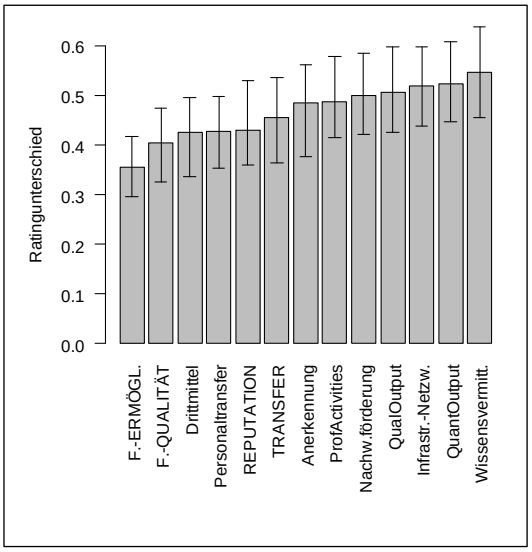

Abbildung 4: Mittlerer Unterschied zwischen den Ratings nach Kategorie

In Bezug auf die Bewertungskriterien Forschungsqualität, Reputation, Forschungsermöglichung und Transfer lässt sich also feststellen, dass hier keine signifikanten Unterschiede zwischen den Unterschieden der beiden jeweiligen Ratings über die vier Kriterien hinweg zu verzeichnen sind. Bei den Bewertungsaspekten sieht das freilich anders aus. Hier sind (wenig überraschend) die Drittmittel besonders klar und eindeutig zu bewerten, die Wissensvermittlung an außeruniversitäre Adressaten eher weniger eindeutig zu bewerten. Vielleicht etwas überraschend ist die Tatsache, dass das Kriterium Forschungsqualität, das auf der vorwiegend qualitativen Beurteilung der eingereichten Schriften und Publikationslisten beruhte, die zweithöchste Übereinstimmung zwischen den beiden Ratern ausweist. Dies zeigt, dass in unserem Fach offenbar relativ klare Standards für die Beurteilung von Publikationen bestehen.

Zusammenfassend lässt sich sagen, dass innerhalb der Gutachtergruppe die in Anschlag zu bringenden Beurteilungskategorien sehr konsistent und über einzelne Bewerter hinweg nachvollziehbar angewendet wurden. Das hier durchgeführte *peer review*-Verfahren hat also insgesamt zu verlässlichen und nachvollziehbaren Ergebnissen geführt.

4. Analyse der Bewertungskategorien

Wir wenden uns nun unter einer anderen Perspektive einer genaueren einrichtungsübergreifenden Analyse der Daten zu. Diese Analyse hat vor allem zum Ziel, Näheres darüber zu erfahren, wie die einzelnen Bewertungskategorien zueinander in Beziehung stehen.

4.1. Bewertungsaspekte

Wir sehen uns zunächst die Bewertungsaspekte an. Datengrundlage hierfür ist wiederum Datensatz A, d.h. die Bewertungen, die vor den Bewertungssitzungen von den Bewertern unabhängig voneinander abgegeben wurden.

Korreliert man die neun Bewertungsaspekte miteinander, so ergeben Korrelationstests (Spearman) für die 36 Korrelationen, dass alle Korrelationen positiv und hoch signifikant sind. Das bedeutet, dass tendenziell höhere Bewertungen für einen gegebenen Bewertungsaspekt mit höheren Bewertungen in jedem der anderen Bewertungsaspekte einhergehen. Allerdings

unterscheiden sich die einzelnen Korrelationen sehr in der Stärke dieses Effekts. Abbildung 5 zeigt die Verteilung der ermittelten Korrelationskoeffizienten.[3]

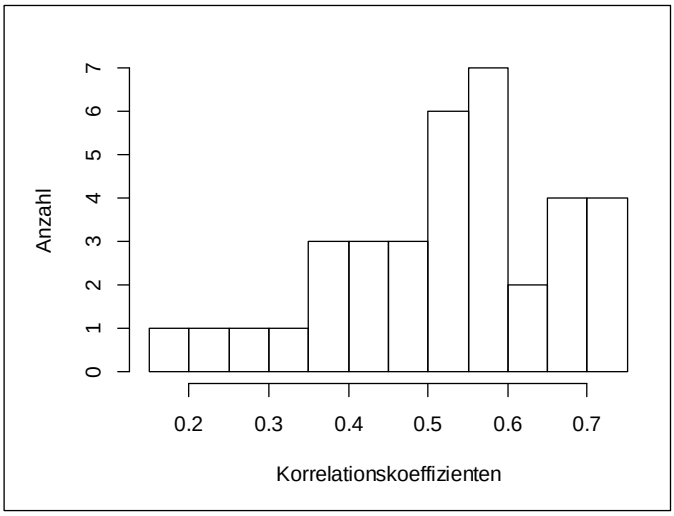

Abbildung 5: Verteilung der Korrelationskoeffizienten für 36 Korrelationen

Eine genauere Analyse dieser Verteilung erweist sich als interessant. Tabelle 4 gibt einen Überblick über die höchsten und niedrigsten Korrelationen.

Korrelation	**Bewertungsaspekt 1**		**Bewertungsaspekt 2**
stark	Qualität des Outputs	mit	Quantität des Outputs
(rho > 0,68)	Professional Activities	mit	Anerkennung
	Professional Activities	mit	Infrastrukturen und Netzwerke
	Drittmittel	mit	Infrastrukturen und Netzwerke
	Personaltransfer	mit	Wissensvermittlung
schwach	Personaltransfer	mit	Qualität des Outputs
(rho ≤ 0,3)	Personaltransfer	mit	Quantität des Outputs
	Wissensvermittlung	mit	Qualität des Outputs

Tabelle 4: Höchste und niedrigste Korrelationen unter den Bewertungsaspekten

3 Korrelationskoeffizienten bewegen sich im Bereich zwischen -1 und +1. Der Wert 0 zeigt die Abwesenheit eines Zusammenhangs an, die Extremwerte perfekte positive bzw. negative Korrelationen.

Wir sehen, dass einige Bewertungsaspekte in enger Beziehung zu anderen stehen. So geht eine hohe Qualität der eingereichten Publikationen einher mit einer insgesamt hohen Zahl an Publikationen, was bedeutet, dass diejenigen, die besonders gute Publikationen vorlegen auch tendenziell diejenigen sind, die mehr publizieren. Andere besonders starke Korrelationen sind vielleicht etwas weniger überraschend. So verwundert es beispielsweise kaum, dass etwa Drittmittel und Infrastrukturaufbau in hohem Maße voneinander abhängen.

Ein besonderes Augenmerk sowohl in der Öffentlichkeit als auch in den universitätsinternen Diskussionen und Verteilungskämpfen liegt bekanntermaßen auf den Drittmitteln. In dieser Diskussion wird oft unterstellt, dass Drittmittel ein Ausweis besonderer Forschungsqualität ist. Die vorliegende Erhebung zeigt aber, dass dies nur bedingt der Fall ist. Zwar gibt es eine positive Korrelation zwischen dem Drittmittelaufkommen und der Qualität und Quantität des Outputs (rho = 0.47, bzw. rho = 0,45), aber diese ist nicht besonders stark, und 70 Prozent der 36 Korrelation sind stärker ausgeprägt.

Abbildung 6 zeigt den Zusammenhang zwischen Drittmitteln und der Qualität des Outputs. Zur besseren Sichtbarmachung der Häufigkeiten sind wieder die Punkte (N = 335) um den eigentlichen Wert herum leicht gestreut. Die durchgezogene schwarze Linie zeigt den Trend in den Daten,4 die gestrichelte Linie stellt eine perfekte Korrelation dar (rho = 1). Abbildung 6 macht deutlich, dass der allgemeine Trend nicht besonders stark ist. So gibt es entlang der x-Achse und an beiden Enden der Drittmittelskala eine breite Streuung in der Qualität des Outputs. Es ist aber auch zu konstatieren, dass Spitzenforschung (mit Ratings von 5 oder 5-4) in aller Regel mit relativ hohen Drittmitteleinwerbungen einhergeht. Umgekehrt lässt sich durch einen Blick auf die rechte Seite des Graphen feststellen, dass hohe Drittmittel nicht automatisch auf herausragende Forschung schließen lassen. Ganz links oben finden wir auch zwei Teilbereiche von Einrichtungen, die ohne bzw. fast ohne Drittmittel herausragende Forschungsqualität erbringen.

4 Die schwarze Linie zeigt das Ergebnis einer nicht-parametrischen Glättungsfunktion (Cleveland 1979).

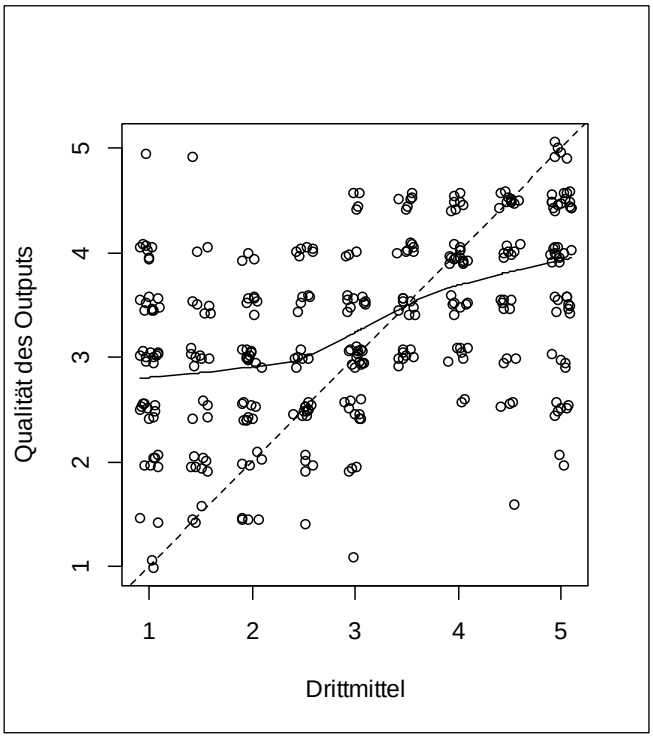

Abbildung 6: Qualität des Outputs nach Drittmitteln

Angesichts dieser Faktenlage erscheint es angebracht, die universitäre Diskussion etwas ehrlicher zu führen und die Drittmittel (zumindest in der Anglistik/Amerikanistik) als das zu behandeln, was sie eigentlich sind: Dringend benötigte Geldmittel für die Forschung an unseren unterfinanzierten Hochschulen, aber kein geeignetes Instrument zur Messung von Forschungsqualtität.

4.2. Bewertungskriterien

Wenden wir uns nun den vier Bewertungskriterien zu. Datengrundlage hierfür ist Datensatz B, d.h. die am Ende beschlossenen Bewertungen. Tabelle 5 ist eine Korrelationsmatrix mit den Korrelationskoeffizienten.

	Reputation	Forschungsermöglichung	Transfer
Forschungsqualität	0.73	0.69	0.39
Reputation	1.00	0.73	0.49
Forschungsermöglichung	0.73	1.00	0.50
Transfer	0.49	0.50	1.00

Tabelle 5: Korrelationskoeffizienten für die Bewertungskriterien

Ein Test der sechs Korrelationen zwischen den vier Bewertungskriterien ergibt, dass alle Bewertungskriterien hoch signifikant positiv miteinander korrelieren ($p = 0$ für alle Korrelation, Spearman). Das Kriterium des Transfers verhält sich aber anders als die drei anderen Kriterien. Während Forschungsqualität, Reputation und Forschungsermöglichung alle sehr stark positiv mit einander korrelieren, steht Transfer nicht in einem so starken Zusammenhang zu den anderen drei Variablen. Dies zeigen auch die Streudiagramme in Abbildung 7. Zur besseren Sichtbarmachung der Häufigkeiten sind wieder die Punkte um den eigentlichen Wert herum leicht gestreut. Die durchgezogene schwarze Linie zeigt den Trend in den Daten, die gestrichelte Linie ist die Diagonale und stellt eine perfekte Korrelation dar ($rho = 1$). Bei den Kriterien Forschungsqualität, Reputation und Forschungsermöglichung sehen wir eine nur mäßige Streuung und einen Trend in den Daten, der sehr nah an der Diagonalen liegt (siehe die Graphen auf der linken Seite von Abbildung 7). Im Gegensatz dazu weicht der Trend bei den drei Graphen auf der rechten Seite von Abb. 7 relativ stark von der Diagonalen ab, mit einer sehr viel größeren Streuung.

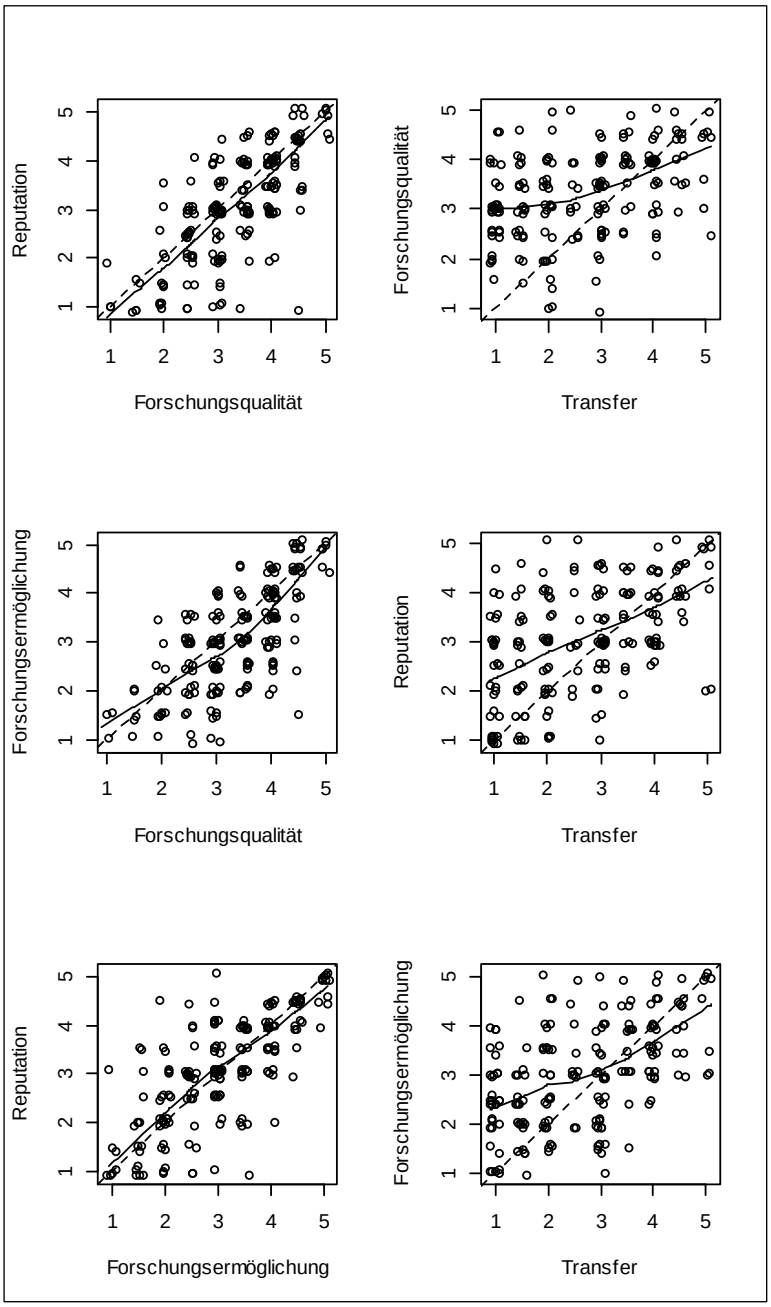

Abbildung 7: Zusammenhänge zwischen den Bewertungskriterien

5. Zusammenfassung und Diskussion

Die vorliegenden statistischen Auswertungen der Ratings haben ergeben, dass die von den Raterpaaren unabhängig voneinander abgegebenen Bewertungen stark übereinstimmen. Dieses Ergebnis kann dahin gehend interpretiert werden, dass die Bewertungskategorien insgesamt relativ gut operationalisiert waren und eine nachvollziehbare, verlässliche Bewertung erlaubten. Es konnte auch gezeigt werden, dass nicht alle Bewertungskriterien gleich eindeutig zu bewerten waren. Insgesamt können wir aber konstatieren, dass in unserem Fach in den einzelnen Teilbereichen klare wissenschaftliche Standards etabliert sind, die eine weitgehend objektive Beurteilung der Forschungsleistungen einer Institution erlauben.

Als Ergebnis der Untersuchung der Zusammenhänge zwischen den einzelnen Bewertungskategorien lassen sich vor allem drei wichtige Ergebnisse festhalten. Erstens korrelieren alle Kategorien mehr oder weniger stark positiv miteinander. Das bedeutet, dass ein gegebener Teilbereich tendenziell ähnliche Bewertungen über alle Kategorien hinweg bekommen hat, was wiederum heißt, dass aus statistischer Sicht die Kategorien zu einem nicht geringen Teil die gleichen zugrundeliegenden Merkmale abprüfen. Dies war vielleicht zum Teil erwartbar, wirft aber auch die Frage auf, inwieweit der große Aufwand, der mit dieser Erhebung verbunden war, gerechtfertigt und notwendig ist. Aus politischer Perspektive ist jedoch zu berücksichtigen, dass die Einbeziehung möglichst vieler Kategorien die Akzeptanz für ein solches Ranking erhöht.

Eine zweite wichtige Erkenntnis ist, dass nicht alle Kategorien gleich gut miteinander korrelieren, und dass insbesondere die Höhe der Drittmittel kein besonders gut geeigneter Indikator für die Forschungsqualität ist. Umgekehrt bedeuten die Ergebnisse, dass in unserem Fach eine qualitative Beurteilung des Forschungsoutputs unerlässlich für eine nachvollziehbare Beurteilung der tatsächlich erbrachten Forschungsleistungen ist.

Drittens haben wir gesehen, dass Transfer in keinem sehr starken Zusammenhang mit den anderen Bewertungskriterien steht. Dies kann dahingehend interpretiert werden, dass in unserem Fach der Transfer an außeruniversitäre Adressaten tendenziell in der Forschung eine weniger prominente Rolle spielt.

Zusammenfassend können wir festhalten, dass die Ergebnisse des Forschungsratings der Anglistik/Amerikanistik als sehr verlässlich und gut nachvollziehbar anzusehen sind. Das wird diejenigen freuen, die gut abgeschnitten haben, und all diejenigen vielleicht betrüben, die ein weniger gutes

Ergebnis erzielt haben. Dies wiederum bringt uns zu der vielleicht entscheidenden Frage, wer denn die Ratingergebnisse nutzen wird, und zu welchem Zweck.

Als primäre Adressaten könnte man an die bewerteten Wissenschaftlerinnen und Wissenschaftler denken, die eine qualifizierte Rückmeldung über die Qualität ihrer Forschungsleistung erhalten. Ob dafür allerdings ein derartiges Rating notwendig ist, darf stark bezweifelt werden, denn die wissenschaftsimmanente Praxis des peer reviews tut dies ohnehin und permanent, so dass jeder Wissenschaftler und jede Wissenschaftlerin durch die ständige Rückmeldung der scientific community sowieso weiß (oder wissen sollte), wo er oder sie qualitätsmäßig steht. Außerdem wurde in dem hier durchgeführten Rating aus datenschutzrechtlichen Gründen gerade darauf verzichtet individuelle Leistungen zu bewerten. Bewertet wurde die Gesamtleistung von Teilbereichen von Institutionen, was notwendigerweise zu Bewertungen führt, die einen wie auch immer gearteten Durchschnittswert wieder geben. Diese Durchschnittswerte wurden von der Gutachtergruppe insbesondere bei unterschiedlichen Leistungsniveaus innerhalb eines Teilbereichs als wenig glücklich empfunden. Auch die Verrechnung von Leistungen ganz unterschiedlicher Professuren (z.B. W3 vs. W2 vs. W1 mit jeweils unterschiedlichen Ausstattungen) stellte sich zuweilen als problematisch dar. Das Rating ist also insgesamt für die bewerteten Einzelpersonen als wenig hilfreich anzusehen, es sei denn, es kann von diesen institutionell zur Verbesserung ihrer Situation genutzt werden.

Dies bringt uns zu weiteren möglichen Adressaten, nämlich die beteiligten Institutionen, die die Ergebnisse als Entscheidungshilfe für Strukturentscheidungen auf Instituts-, Fakultäts- und Universitätsebene nutzen könnten. Wie dies im Einzelnen aussehen mag, bleibt dahingestellt, aber generell ist wohl zu befürworten, dass institutionelle Entscheidungen besser auf einer verlässlichen Datengrundlage beruhen sollten als auf Hörensagen oder Einflüsterungen einflussreicher Personen. Das vom Wissenschaftsrat durchgeführte Rating stellt eine solche Datengrundlage in Bezug auf die Forschungsleistung bereit. Dass Forschungsleistungen dabei nur ein Kriterium in einer komplexen institutionellen Gemengelage darstellt, versteht sich von selbst, und unsere Ergebnisse zeigen, dass die strukturellen Verhältnisse an vielen der untersuchten Universitäten, Pädagogischen Hochschulen und Instituten die Forschungsleistung eher beeinträchtigen als befördern.

Eingang des revidierten Manuskripts 30.12.2012

Literaturverzeichnis

Cleveland, William S. (1979), Robust locally weighted regression and smoothing scatterplots. In: *Journal of the American Statistical Association* 74, 829-836.

Gries, Stefan (2008), *Statistik für Sprachwissenschaftler*. Göttingen: Vandenhoeck & Ruprecht.

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

Wissenschaftsrat (2010), *Empfehlungen zur vergleichenden Forschungsbewertung in den Geisteswissenschaften* (Drs. 10039-10). Köln. www.forschungsrating.de

Wissenschaftsrat (2012a), *Ergebnisse des Forschungsratings Anglistik und Amerikanistik* (Drs. 2756-12). Köln. www.forschungsrating.de

Wissenschaftsrat (2012b), *Hintergrundinformation: Pilotstudie Forschungsrating im Fach Anglistik und Amerikanistik*. Köln. www.forschungsrating.de