WISSENSCHAFTSRAT

GESCHÄFTSSTELLE

Drs. 8453-08
Köln, 10.04.2008

# Bericht der Steuerungsgruppe zur Pilotstudie Forschungsrating Chemie und Soziologie

Inhalt Seite

## Vorbemerkung

Der Wissenschaftsrat hat im November 2004 Empfehlungen zu Rankings im Wissenschaftssystem vorgelegt. Darin hat er bestehende Rankings einer methodischen Kritik unterzogen und ein Verfahren für ein Forschungsrating entwickelt.[1] Um dieses Verfahren zu erproben, hat der Wissenschaftsrat im Juli 2005 beschlossen, eine Pilotstudie in den Fächern Chemie und Soziologie durchzuführen. Mit ihrer Durchführung hat er eine Steuerungsgruppe beauftragt, der neben Mitgliedern des Wissenschaftsrats auch Vertreter der Wissenschaftsorganisationen und weitere Sachverständige angehört haben. Diese Steuerungsgruppe hat ihrerseits für jedes der beiden Fächer der Pilotstudie eine Bewertungsgruppe eingesetzt, die aus nationalen und internationalen Fachgutachtern bestand.

Der vorliegende Bericht enthält eine Beschreibung und Bewertung des Verlaufs der Pilotstudie Forschungsrating in den Fächern Chemie und Soziologie durch die Steuerungsgruppe sowie Empfehlungen zur Zukunft des Forschungsratings. Er fußt wesentlich auf den Erfahrungen der beiden Bewertungsgruppen, die von diesen in Abschlussberichten[2] festgehalten wurden. Die Abschlussberichte sind dem Bericht als Anlagen beigefügt und wurden von der Steuerungsgruppe nicht verändert.

Die Steuerungsgruppe hat den Bericht zur Pilotstudie Forschungsrating Chemie und Soziologie am 10. April 2008 verabschiedet.

---

1 Wissenschaftsrat: Empfehlungen zu Rankings im Wissenschaftssystem. Teil 1: Forschung, in: Empfehlungen und Stellungnahmen 2004, Köln 2005, S. 159 – 220.

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

## Kurzfassung

Der Wissenschaftsrat hat im November 2004 das Konzept für ein Verfahren zu einem Forschungsrating veröffentlicht und empfohlen, dieses Verfahren in einer Pilotstudie zu erproben. Diese Pilotstudie wurde nach einem Beschluss vom Juli 2005 in den Fächern Chemie und Soziologie durchgeführt. Sie begann im Herbst 2005 und wurde Anfang 2008 abgeschlossen.

Mit der Durchführung der Pilotstudie beauftragte der Wissenschaftsrat eine Steuerungsgruppe, die ihrerseits für jedes der bearbeiteten Fächer aus Fachgutachtern bestehende Bewertungsgruppen einsetzte. Der vorliegende Bericht fasst die Erfahrungen der Steuerungsgruppe aus der Pilotstudie zusammen, die sich bei ihren Beratungen unter anderem auf die in der Anlage veröffentlichten Abschlussberichte der beiden Bewertungsgruppen gestützt hat.

Das Forschungsrating zeichnet sich gegenüber herkömmlichen Rankings durch eine Reihe von Alleinstellungsmerkmalen aus:

- Die Forschungsqualität wird durch „Informed Peer Review“ auf Basis quantitativer und qualitativer Vergleichsdaten differenziert und unter Berücksichtigung von Kontextinformationen vergleichend bewertet;
- die Fachgemeinschaften wirken an der Definition und Operationalisierung der Bewertungskriterien mit;
- durch den Ausweis der Differenzierung der Forschungsqualität innerhalb der einzelnen Einrichtungen haben die Ergebnisse einen höheren Informationsgehalt;
- durch die Bewertung nach mehreren Kriterien spiegeln sich unterschiedliche Leistungsprofile der Einrichtungen in den Ergebnissen wider;
- durch die Einbeziehung der außeruniversitären Forschungseinrichtungen, die in Fächern wie der Chemie eine große Rolle spielen, wird ein umfassendes Bild der deutschen Forschungslandschaft gezeichnet.

Auch wenn die Forschungsqualität als zentrales Kriterium gilt, hängt die Relevanz der Bewertung nach den einzelnen Kriterien Forschungsqualität, Impact/Effektivität, Effizienz, Nachwuchsförderung, Transfer in andere gesellschaftliche Bereiche und Wissensvermittlung/Wissensverbreitung für eine bestimmte Einrichtung auch von

deren Aufgabenstellung ab. Diese muss deshalb bei der Interpretation der Bewertungen berücksichtigt werden.

Die Pilotstudie hat gezeigt, dass das Forschungsrating an Fächer mit sehr unterschiedlicher Forschungspraxis angepasst werden kann. So beruht beispielsweise die Bewertung des zentralen Kriteriums „Forschungsqualität“ in der Chemie unter anderem auch auf Zitationsindikatoren, während in der Soziologie auf solche Indikatoren aufgrund der heterogenen Publikationspraxis und der daraus resultierenden unzureichenden Datenlage verzichtet werden musste. Stattdessen beruht die Bewertung der Forschungsqualität in der Soziologie zu einem erheblichen Teil auf der Lektüre ausgewählter Publikationen durch die Gutachter. Beide Vorgehensweisen haben zu differenzierten und, wie an der hohen Gutacherübereinstimmung zu erkennen ist, auch verlässlichen Bewertungen geführt.

Verlauf und Ergebnisse der Pilotstudie berechtigen zu der Annahme, dass das Forschungsrating auch in weiteren Fächern mit Erfolg angewandt werden kann. Aus Sicht der Steuerungsgruppe sollte das Verfahren schrittweise weiterentwickelt werden. Im Zuge dessen sollte die Definition und Datengrundlage einzelner Kriterien weiter verbessert werden. Insbesondere sollte geprüft werden, ob Unterschiede in der Belastung durch Aufgaben neben der Forschung, beispielsweise durch die Lehre, Unterschiede der verfügbaren Forschungsinfrastruktur sowie der in verschiedenen Teilgebieten der jeweiligen Fächer unterschiedliche Ressourcenbedarf bei der Effizienzbewertung mit vertretbarem Aufwand detaillierter berücksichtigt werden kann. Ein Desiderat bleibt auch die weitere, fachspezifische Präzisierung der Kriterien in der Dimension Wissenstransfer und die Verbesserung der Datengrundlage dazu.

Der Aufwand für die Bewertung der Chemie und der Soziologie war erheblich, für eine Pilotstudie aber angemessen. Bei einer Weiterentwicklung des Forschungsratings muss der Aufwand in vertretbaren Grenzen gehalten und die Aussagekraft der Ergebnisse optimiert werden. Die Steuerungsgruppe empfiehlt:

- Die Definition der Forschungseinheiten stärker zu vereinheitlichen;
- den Umfang der erhobenen Daten zu reduzieren;
- die Vorlaufzeit für die Datenerhebung zu vergrößern, um den bewerteten Einrichtungen die Vorbereitung zu erleichtern;

- die Erhebungsformate gemeinsam mit anderen Daten erhebenden Einrichtungen nach Möglichkeit so zu standardisieren, dass Daten mehrfach nutzbar sind; sowie
- die Datenqualität weiter zu optimieren, um die Gutachter zu entlasten.

Wenn das Verfahren für ein Forschungsrating in diesem Sinne weiterentwickelt werden sollte, empfiehlt die Steuerungsgruppe, als nächstes je ein Fach aus den Geistes- und den Technikwissenschaften zu bewerten, da diese sich sowohl hinsichtlich der fachinternen Publikations- und Kommunikationswege als auch hinsichtlich der für die Transferdimension relevanten Beziehungen zu anderen gesellschaftlichen Bereichen am stärksten von den Natur- und Sozialwissenschaften unterscheiden.

Vor einer Entscheidung über die dauerhafte Einführung des Forschungsratings sollte der Nutzen der veröffentlichten Ergebnisse der Pilotstudie, auch im Vergleich mit eventuell vorhandenen eigenen Evaluationen, in einem Dialog mit den Adressaten weiter geklärt werden. Zudem sollte der Wissenschaftsrat die Untersuchung der Folgen dieses und ähnlicher Bewertungsverfahren unterstützen.

# A. Ausgangslage

## A.I. Wissenschaftspolitischer Kontext

Das System der wissenschaftlichen Institutionen in Deutschland hat sich in den letzten Jahren deutlich verändert. So ist heute anerkannt, dass das Hochschulsystem nur dann sowohl den wachsenden Bedarf an tertiärer Bildung decken als auch international herausragende Forschungsleistungen erbringen und die Unternehmen durch praxisorientierte Forschung und Weiterbildungsangebote unterstützen kann, wenn die einzelnen Hochschulen Schwerpunkte bilden und sich unterschiedliche Hochschulprofile im Wettbewerb herausdifferenzieren.[3] Diese Differenzierung wird inzwischen von Bund und Ländern – paradigmatisch etwa in der Exzellenzinitiative – aktiv gefördert. Zugleich werden starre Abgrenzungen der verschiedenen Sektoren von universitärer und außeruniversitärer Forschung vor dem Hintergrund dieser wettbewerblichen Differenzierung durch zahlreiche Kooperationen, die an Häufigkeit, Intensität und Verbindlichkeit in den letzten Jahren deutlich zunehmen, mehr und mehr aufgelöst.

Ein wichtiger Faktor in diesen Veränderungen ist die Umstellung des Verhältnisses zwischen Staat und Wissenschaft von einem klassisch-bürokratischen zu einem stärker wettbewerblichen System. Von Zielvereinbarungen erhofft man sich, dass sie an die Stelle staatlicher Regulierung und ministerialer Entscheidungen treten, wobei die wissenschaftlichen Institutionen in unterschiedlichem Umfang größere Spielräume erhalten sollen, die vereinbarten Ziele in weitgehend autonomer Weise zu erreichen. Die Anforderungen an die Selbststeuerungsfähigkeiten der Einrichtungen und damit an strategisches Steuerungswissen sind deshalb erheblich gestiegen. Gleichzeitig erzeugt die wettbewerbliche Differenzierung auch einen größeren Bedarf an Orientierungswissen für Studienanfänger, wissenschaftlichen Nachwuchs und Kooperationspartner wissenschaftlicher Einrichtungen. Nicht zuletzt zieht die zunehmende Autonomie wissenschaftlicher Einrichtungen die Forderung von Politik und Gesellschaft nach mehr Transparenz über die erbrachten Leistungen nach sich.

Vor diesem Hintergrund haben öffentliche Leistungsvergleiche wissenschaftlicher Einrichtungen in Gestalt von Rankings, wie sie für den Hochschulsektor seit Jahren regelmäßig von großen Publikumszeitschriften publiziert werden, erheblich an Be-

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

deutung gewonnen. Ihre Wirkung beschränkt sich nicht auf die publikumswirksame Verkündung von ersten und letzten Plätzen. Rankings, die die Bedingungen des Studiums an verschiedenen Hochschulen zum Gegenstand haben, haben zum Ziel, potentielle Studierende bei der Entscheidung für Studienort und –fach zu unterstützen, und nehmen damit Einfluss auf die Rekrutierungschancen der Einrichtungen. Internationale Rankings – vor allem das Ranking des Times Higher Education Supplement und das sogenannte Shanghai-Ranking[4] – haben großen Einfluss sowohl auf die Formulierung globaler wissenschaftspolitischer Ziele als auch in den strategischen Überlegungen der einzelnen Einrichtungen.

Angesichts der Folgen, die Rankings für die wissenschaftlichen Institutionen haben, und ihrer häufig unkritischen Aufnahme ist es bedenklich, dass die Methodik vieler Zeitschriftenrankings nicht transparent ist und die Wissenschaft bei ihrer Fortentwicklung keine Mitsprache hat. Im Bereich der Forschung ist zudem problematisch, dass die internationalen Rankings den in Deutschland in vielen Fachgebieten besonders wichtigen außeruniversitären Bereich nicht erfassen. Der Wissenschaftsrat hat sich deshalb im Jahr 2004 mit der Funktion und Methodik von vergleichenden Leistungsbewertungen in der Wissenschaft befasst, Standards für solche Verfahren formuliert und einen Vorschlag für ein Verfahren zur Bewertung der Forschungsleistungen von Universitäten und außeruniversitären Forschungseinrichtungen vorgelegt.[5]

Bei der Entwicklung seines Verfahrensvorschlags ist der Wissenschaftsrat davon ausgegangen, dass die vorhandenen Rankings weiter bestehen und die Entwicklung des Wissenschaftssystems in erheblichem Maße beeinflussen werden. Ihre Auswirkungen werden sich deshalb nur begrenzen lassen, indem den Rankings der Publikumszeitschriften ein differenzierteres, wissenschaftlich geleitetes Verfahren entgegengesetzt wird. Unerwünschte Anreizeffekte von Leistungsvergleichen sind dabei zu beobachten und ggf. durch eine differenzierte Kommunikation der Ergebnisse, durch eine kontinuierliche Weiterentwicklung der Verfahren und durch begleitende Maßnahmen auszugleichen.

---

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

5 Wissenschaftsrat: Empfehlungen zu Rankings im Wissenschaftssystem. Teil 1: Forschung, in: Empfehlungen und Stellungnahmen 2004, Köln 2005, S. 159 – 220.

Der vom Wissenschaftsrat vorgelegte Verfahrensvorschlag sieht eine Bewertung der Leistungen von Universitäten und außeruniversitären Forschungseinrichtungen in den Dimensionen Forschung, Nachwuchsförderung und Wissenstransfer vor, klammert hingegen die Lehre aus. Zwar ist dem Wissenschaftsrat bewusst, dass die Lehre eine ebenso fundamentale Aufgabe der Hochschulen und die Verbesserung ihrer Qualität ein dringendes Desiderat ist. Während jedoch Standards für eine hochschulübergreifende, vergleichende Bewertung von Forschungsleistungen international anerkannt und auch theoretisch gut begründet sind, ist dies für die Lehre nicht der Fall. Verfahren, die beides gleichzeitig bewerten, sind aus methodischen Gründen nicht sinnvoll; für ein solches integriertes Verfahren gibt es auch international kein anerkanntes Vorbild. Leistungsvergleiche der Lehre sollten in eine systematische Initiative zur Verbesserung der Qualität von Studium und Lehre eingebettet werden und müssen auf den Anstrengungen der einzelnen Hochschulen zur Verbesserung ihres Qualitätsmanagements basieren. Sie sind deshalb erst als ein späterer Schritt sinnvoll.[6]

## A.II. Die Pilotstudie Forschungsrating

### II.1. Vorgeschichte und Beschlusslage im Wissenschaftsrat

In seinen Empfehlungen zu Rankings im Wissenschaftssystem vom November 2004[7] hat der Wissenschaftsrat die Grundzüge eines fächerspezifischen, mehrdimensionalen Forschungsratings entwickelt. Das Forschungsrating legt das Prinzip „Informed Peer Review“ zugrunde. Anders als bei den bekannten Rankings werden also keine Rangplätze berechnet, sondern Bewertungen durch Gutachter vorgenommen, die dazu standardisierte und statistisch ausgewertete Daten über die einzelnen Einrichtungen erhalten. Ziel ist, den Wettbewerb in der Forschung effektiver und effizienter zu machen, indem die Transparenz der Forschungsleistungen im öffentlichen Sektor erhöht wird, und die forschenden Einrichtungen bei ihrer Profilierung im Rahmen ihrer jeweiligen Mission zu unterstützen, indem ihnen durch vergleichende Leistungsbewertungen eine Standortbestimmung nach international gültigen Maßstäben ermöglicht wird. Der Wissenschaftsrat geht davon aus, dass die Ergebnisse eines Forschungsratings außerdem für akademische und nicht-akademische

---

6 Der Wissenschaftsrat wird auf seinen Frühjahrssitzungen 2008 über Empfehlungen zur Qualität der Lehre beraten.

7 Wissenschaftsrat 2005.

Kooperationspartner aus dem In- und Ausland sowie für Wissenschaftler, insbesondere für Nachwuchswissenschaftler, von erheblichem Interesse sind.

Der Wissenschaftsrat hat in seinen Empfehlungen die Grundzüge des Forschungsratings entwickelt, jedoch festgehalten, dass die Eignung der Methode vor einer Entscheidung über die Einführung eines regelmäßigen, alle Fächer abdeckenden Verfahrens in einer Pilotstudie erprobt werden müsse. Deshalb hat er eine Steuerungsgruppe eingesetzt, die neben Mitgliedern seiner wissenschaftlichen Kommission und weiteren Sachverständigen auch institutionelle Vertreter der Deutschen Forschungsgemeinschaft (DFG), der Fraunhofer-Gesellschaft (FhG), der Helmholtz-Gemeinschaft (HGF), der Hochschulrektorenkonferenz (HRK), der Max-Planck-Gesellschaft (MPG) und der Leibniz-Gemeinschaft (WGL) auf Vizepräsidenten-Ebene sowie Gäste aus Länderministerien und dem BMBF umfasst. Er beauftragte diese zunächst mit der Vorbereitung und im Juli 2005 mit der Durchführung einer solchen Pilotstudie. Als eines der Fächer für die Pilotstudie wurde auch auf Vorschlag der Gesellschaft Deutscher Chemiker (GDCh) sowie des Verbands der Chemischen Industrie (VCI), die sich zu einer auch finanziellen Unterstützung der Pilotstudie bereit erklärten, die Chemie ausgewählt. Als zweites, methodisch von der Chemie deutlich unterschiedenes Fach wählte der Wissenschaftsrat die Soziologie. Die Pilotstudie begann im Oktober 2005.

Über den Verlauf der Pilotstudie wurde dem Wissenschaftsrat im Mai 2007 ein Zwischenbericht erstattet. Im Dezember 2007 hat die Steuerungsgruppe die Ergebnisse der Bewertung der Chemie, im April 2008 die Ergebnisse der Bewertung der Soziologie veröffentlicht.[8]

## II.2. Organisation und Ablauf der Pilotstudie

Um das Verfahren an die beiden für die Pilotstudie ausgewählten Fächer anzupassen und die Bewertungen durchzuführen, setzte die Steuerungsgruppe zwei Bewertungsgruppen ein, die aus 15 respektive 16 Gutachtern[9] bestanden. Gutachtervorschläge wurden von DFG, FhG, HGF, HRK, MPG und WGL sowie für die Chemie

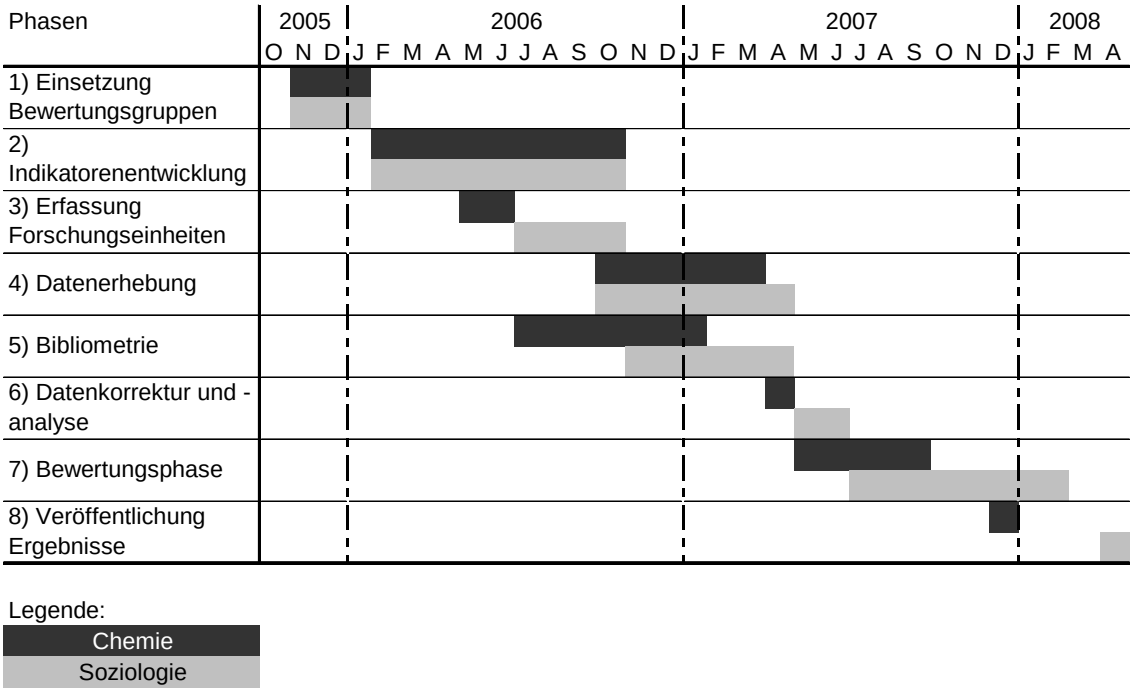
[8] Steuerungsgruppe der Pilotstudie Forschungsrating im Auftrag des Wissenschaftsrates: Forschungsleistungen deutscher Universitäten und außeruniversitärer Einrichtungen in der Chemie. Köln, 18.12.2007; dies.: Forschungsleistungen deutscher Universitäten und außeruniversitärer Einrichtungen in der Soziologie. Köln, 10.04.2008.

[9] Aus Gründen der Lesbarkeit sind hier und im Folgenden nicht die männliche und weibliche Sprachform nebeneinander aufgeführt. Personenbezogene Aussagen, Amts-, Status-, Funktions- und Berufsbezeichnungen gelten aber stets für Frauen und für Männer.

von GDCh und VCI, für die Soziologie von der Deutschen Gesellschaft für Soziologie (DGS) eingeholt. Bei der Besetzung der Bewertungsgruppen achtete die Steuerungsgruppe auf eine breite Abdeckung der wichtigsten Teilgebiete des jeweiligen Fachs und auf die internationale Erfahrung der Experten. Um die internationale Perspektive auch formal einzubinden, wurden auch Gutachter aus den Niederlanden, Österreich und der Schweiz berufen. In beide Gruppen wurden neben akademischen Gutachtern jeweils zwei Vertreter einer fachnahen Praxis berufen. Im Fall der Chemie handelt es sich um Wissenschaftler aus forschenden Industrieunternehmen, in der Soziologie um Vertreter eines Umfrageinstituts sowie einer Stiftung. Schließlich wurde je ein Vertreter der Steuerungsgruppe als Berichterstatter in jede Bewertungsgruppe entsandt. In der Geschäftsstelle des Wissenschaftsrats wurde zur Betreuung der Pilotstudie ein Projektteam eingerichtet.

Die wichtigsten Phasen der Pilotstudie und der Zeitbedarf dieser Phasen in den beiden Fächern sind dem folgenden Schema zu entnehmen.

**Abbildung 1: Ablauf- und Zeitplan der Pilotstudie Forschungsrating**




Die erste Aufgabe der Bewertungsgruppen (Phase 2) bestand darin, die für ihr jeweiliges Fach angemessenen quantitativen und qualitativen Indikatoren für die Bewertung zu bestimmen. Als Rahmenbedingung war aufgrund eines Vorschlags der Steuerungsgruppe im Wissenschaftsrat beschlossen worden, die Forschungsqualität als

das zentrale Kriterium des Forschungsratings differenzierter zu bewerten als die anderen Kriterien. Dazu wurde die Ebene sogenannter „Forschungseinheiten“ definiert, die bei den Universitäten in der Regel unterhalb der Fachbereichsebene, bei den außeruniversitären Einrichtungen auf Abteilungsebene angesiedelt waren. Ziel dieser Phase war die jeweils fachspezifische Zuordnung von Indikatoren zu Kriterien in einer sogenannten Bewertungsmatrix. Der erste Entwurf dazu wurde in einem Pretest, an dem insgesamt acht Einrichtungen teilnahmen, erprobt und daraufhin korrigiert und vereinfacht. Die endgültige Bewertungsmatrix wurde veröffentlicht und lag dem weiteren Erhebungs- und Bewertungsprozess zugrunde.[10] Die ursprünglich neun vom Wissenschaftsrat empfohlenen Kriterien wurden von den Bewertungsgruppen in Absprache mit der Steuerungsgruppe auf sechs reduziert, um das Verfahren besser handhabbar zu gestalten.

**Abbildung 2: Raster der Dimensionen und Kriterien nach Vereinfachung durch die Bewertungsgruppen**

| Dimension | Kriterium |
|---|---|
| Forschung | I. Forschungsqualität (Ebene Forschungseinheit) |
| | II. Impact/Effektivität |
| | III. Effizienz |
| Nachwuchsförderung | IV. Nachwuchsförderung |
| Wissenstransfer | V. Transfer in andere gesellschaftliche Bereiche |
| | VI. Wissensvermittlung und -verbreitung |

In diese Arbeitsphase gehörte auch, das jeweilige Fach durch Bestimmung der zugehörigen Teilgebiete genauer abzugrenzen und die Stufen der Bewertungsskala zu definieren sowie die Fragebögen für die Datenerhebung zu entwickeln.

Voraussetzung für die Datenerhebung war, die zu bewertenden Forschungseinheiten und die ihnen zugeordneten Wissenschaftler – in der Chemie: die leitenden Wissenschaftler – bei den teilnehmenden Universitäten und außeruniversitären Forschungseinrichtungen zu erfassen (Phase 3). Zu diesem Zweck wurden die Einrichtungen

[10] Zu Details der Bewertungsmatrizes und den einzelnen Indikatoren vgl. die Abschlussberichte der beiden Bewertungsgruppen, Pilotstudie Forschungsrating Chemie: Abschlussbericht der Bewertungsgruppe (Wissenschaftsrat, Drs. 8370-08) und Pilotstudie Forschungsrating Soziologie: Abschlussbericht der Bewertungsgruppe (Wissenschaftsrat, Drs. 8422-08).

zunächst gebeten, je Fach einen Wissenschaftler als sogenannten Fachkoordinator zu benennen. Wegen des engen Zeitplans der Pilotstudie wurde die Benennung der Fachkoordinatoren und die Erfassung der Forschungseinheiten parallel zur Indikatorenentwicklung durch die Gutachter vorgenommen.

Für die Datenerhebung (Phase 4) wurden den Fachkoordinatoren Fragebögen und vorformatierte Tabellen übersandt, die per e-mail versandt und mit handelsüblichen Office-Programmen bearbeitet werden konnten. Je Einrichtung und Fach war ein übergreifender Fragebogen und ein Fragebogen pro Forschungseinheit samt Tabellenanhang auszufüllen. In der Soziologie, für die keine repräsentativen Zitationsdaten vorliegen, wurden die Forschungseinheiten zudem gebeten, eine festgelegte Anzahl an exemplarischen Publikationen in elektronischer Form einzureichen. Für alle Daten und sonstigen Angaben wurde ein fünfjähriger Erhebungszeitraum vom 1.1.2001 bis 31.12.2005 festegelegt.

Sobald die Forschungseinheiten erfasst und die zugehörigen Wissenschaftler benannt waren – für die Chemie war dies schon vor Abschluss der Fragebogenentwicklung der Fall – , wurden das Institut für Wissenschafts- und Technikforschung der Universität Bielefeld für die Chemie und das Informationszentrum Sozialwissenschaften der GESIS in Bonn für die Soziologie beauftragt, die Publikationen der gemeldeten Wissenschaftler aus dem Erhebungszeitraum für eine Publikations- und, im Fall der Chemie, auch eine Zitationsanalyse zu erfassen (Phase 5). Für die Chemie wurde dafür das Web of Science von Thomson Scientific verwendet. Das IZ Sozialwissenschaften benutzte eine Kombination aus seiner eigenen Datenbank SOLIS und verschiedenen Datenbanken von Cambridge Scientific Abstracts. Beide Institutionen stellten die Rechercheergebnisse auf passwortgeschützte Internetseiten und baten die bewerteten Einrichtungen, die Publikationslisten zu überprüfen und ggf. zu korrigieren.

Alle Daten wurden von der Geschäftsstelle auf Einhaltung der Erhebungsregeln, Konsistenz und Plausibilität überprüft und Korrekturen mit den Fachkoordinatoren abgestimmt (Phase 6). Anschließend fasste die Geschäftsstelle die Erhebungsergebnisse mit weiteren, von externen Kooperationspartnern bezogenen Daten zu einem Datenbericht je Einrichtung und Fach zusammen, der der Einrichtung zur Abschlusskontrolle noch einmal übersandt wurde, und nahm dann statistische Analysen

vor. Zentral war insbesondere die Berechnung von Perzentilwerten[11] für alle quantitativen Indikatoren, die es den Gutachtern ermöglichten, die relative Position der jeweiligen Einrichtung im Fach unabhängig von der Messeinheit auf einen Blick zu erfassen.

Auf Basis der Datenberichte, der Publikationslisten der Forschungseinheiten und in der Soziologie auch der eingereichten Publikationen nahmen die Gutachter die Bewertung der Einrichtungen und Forschungseinheiten nach den einzelnen Kriterien vor (Phase 7). Dabei wurden zunächst jeweils zwei Gutachter als Berichterstatter gebeten, unabhängig voneinander zu einem Bewertungsvorschlag zu kommen. Falls Dissens bestand, wurde den Berichterstattern Gelegenheit gegeben, die Gründe für diesen Dissens zu klären. Anschließend wurden alle Noten von den Bewertungsgruppen plenar diskutiert und in einem weiteren Durchgang mit Blick auf die Gesamtverteilung endgültig abgestimmt.

## II.3. Erfahrungen aus der Datenerhebung[12]

Auf die im April 2006 an alle staatlichen sowie ausgewählte nicht-staatliche Universitäten und an die Trägerorganisationen der außeruniversitären Forschung versandte Anfrage, ob sie am Forschungsrating Chemie und/oder Soziologie teilnehmen, antworteten insgesamt 78 Einrichtungen für die Chemie und 64 Einrichtungen für die Soziologie positiv. Eine Reihe von Einrichtungen zog aus unterschiedlichen Gründen (s. u., S. 17) die Teilnahme im Verlauf der Studie zurück, so dass letztlich 57 Universitäten und 20 außeruniversitäre Institute in der Chemie sowie 54 Universitäten und 3 außeruniversitäre Institute in der Soziologie teilnahmen. Zum Stichtag 31.12.2005 umfassten die gemeldeten Forschungseinheiten 1038 (Chemie) respektive 376 (Soziologie) besetzte Professuren. Für die Chemie liegt diese Zahl deutlich über der vom Statistischen Bundesamt angegebenen Zahl der Universitätsprofessor/-innen im Lehr- und Forschungsbereich „Chemie“ von 895 im Jahr 2005[13]; für die Soziologie, wo aufgrund der unterschiedlichen Fachklassifikation keine Vergleichszahlen der amtlichen Statistik vorliegen, ist nach einem Vergleich mit der Gesamterhebung der

---

[11] Der Perzentilwert gibt an, wie viele von allen bewerteten Einheiten in Bezug auf einen bestimmten Indikator höchstens den gleichen Wert erreichen wie die betrachtete Einheit. Bspw. bedeutet ein Perzentilwert von 90 % bezogen auf die Publikationszahl einer Einrichtung, dass 90 % aller Einrichtungen weniger als oder höchstens genauso viel wie die betrachtete Einrichtung publiziert haben. 0 % bedeutet immer den geringsten, 100 % den höchsten gemessenen Wert. Der Perzentilwert von 50 % heißt auch „Median“.

[12] Die in diesem und den folgenden Abschnitten zusammengefassten Erfahrungen aus der Pilotstudie sind den Abschlussberichten der beiden Bewertungsgruppen entnommen und dort ausführlicher beschrieben und bewertet.

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

Deutschen Gesellschaft für Soziologie aus dem Jahr 2004 mit Ausnahme der erwähnten Rückzieher ebenfalls von einer flächendeckenden Erfassung auszugehen.

Die Datenerhebung erforderte zum Teil einen erheblichen Einsatz der Fachkoordinatoren, da die benötigten Daten vor allem an den Universitäten häufig nicht zentral vorlagen, sondern bei einzelnen Instituten, bisweilen auch bei einzelnen Wissenschaftlern abgefragt werden mussten. Auch die Unterstützung durch die Verwaltung, die bspw. Drittmitteldaten liefern sollte, lief nicht überall reibungslos. Hinzu kam, dass die elektronischen Fragebögen, die, um Einheitlichkeit und Verarbeitbarkeit zu sichern, Vorformatierungen sowie Längenbeschränkungen aufwiesen, aus Sicht vieler Fachkoordinatoren zu umständlich zu handhaben waren. Vielfach wurde der Wunsch nach einer Online-Erhebung geäußert.

Für die Datenerhebung wurde der Zeitraum 2001 – 2005 mit Stichtag 31.12.2005 festgelegt. Für die Chemie wurde dabei das sogenannte „Work Done At“-Prinzip befolgt, d. h., alle in diesem Zeitraum an der jeweiligen Einrichtung beschäftigten Wissenschaftler wurden mit ihren in diesem Zeitraum und an der Einrichtung erbrachten Leistungen dieser Einrichtung zugerechnet, auch wenn sie zum Ende des Erhebungszeitraums hin an eine andere Einrichtung gewechselt oder in den Ruhestand gegangen waren. Demgegenüber wurde in der Soziologie das „Current Potential“-Prinzip befolgt, wonach die am Stichtag an der Einrichtung beschäftigten Wissenschaftler der Einrichtung angerechnet werden, diese aber mit allen im Erhebungszeitraum erbrachten Leistungen.

Das „Current Potential“-Prinzip hat zur Folge, dass Stellen, die am Stichtag nicht besetzt sind, nicht in die Wertung eingehen. Dies erlaubte einerseits eine aktuellere Einschätzung der Leistungsfähigkeit auch mit Blick auf künftig erwartbare Leistungen, trug aber andererseits mit dazu bei, dass in der Soziologie mit 7 % deutlich mehr Forschungseinheiten nicht bewertet werden konnten als in der Chemie (2 %), und war auch einer der Gründe dafür, dass in der Soziologie mit sieben Universitäten mehr Einrichtungen als in der Chemie (eine Universität) nach Bekanntgabe der Erhebungsmodalitäten ihre Teilnahme zurückzogen.

Die für die differenzierte Bewertung der Forschungsqualität gedachten, unterhalb der Ebene der Einrichtungen angesiedelten Forschungseinheiten wurden von den Bewertungsgruppen entsprechend den Kooperationsgepflogenheiten in ihren Fächern

unterschiedlich definiert. Im Resultat wurden in der Chemie in der Regel größere Einheiten mit durchschnittlich sechs leitenden Wissenschaftlern, davon drei Professoren, gemeldet, während in der Soziologie fast 75 % aller Einheiten nur je eine Professur umfassten. Schwierigkeiten bereitete den Fachkoordinatoren bisweilen die Zuordnung von interdisziplinären Einheiten. Diese waren in der Chemie in der Minderheit, während in der Soziologie die Mehrzahl der Forschungseinheiten nach eigener Einschätzung interdisziplinär forscht. Trotzdem wurde die Interdisziplinarität in der Chemie häufiger problematisiert, wobei vor allem auf den „artifiziellen“ Charakter der zu bildenden Forschungseinheiten sowie auf das Problem hingewiesen wurde, interdisziplinäre Einheiten mit solchen im Kern des jeweiligen Fachgebiets zu vergleichen.

Die für den Umgang mit gemeinsamen Berufungen gedachte Option, einrichtungsübergreifende Forschungseinheiten anzumelden, wurde kaum genutzt. Weitere, fachspezifische Besonderheiten der Forschungseinheitenstruktur (Serviceeinheiten; An-Institute) werden in den Abschlussberichten der Bewertungsgruppen diskutiert.

Die enge Zeitplanung der Pilotstudie brachte es mit sich, dass der Stichtag und damit auch der Erhebungszeitraum zum Zeitpunkt des Versands der Fragebögen bereits in der Vergangenheit lag. Die dadurch notwendige retrospektive Erhebung bereitete vielen Universitäten, die mit vergleichbaren Datenanforderungen bis dato nicht konfrontiert gewesen waren, erhebliche Probleme und wurde auch häufig als Grund dafür angeführt, dass die ursprünglich vorgesehene Erhebungsfrist von zwei Monaten zu kurz sei. Sie wurde auf Wunsch vieler Fachkoordinatoren deshalb um sechs Wochen verlängert. Viele Fachkoordinatoren rieten darüber hinaus, die genauen Datenanforderungen künftig mit einer erheblichen Vorlauffrist vor dem Stichtag bekannt zu geben, um eine Datensammlung im laufenden Geschäft zu ermöglichen.

Die Datenüberprüfung in der Geschäftsstelle ergab in beiden Fächern bei je 85 – 90 % aller Einrichtungen einen Korrekturbedarf, der eine Rücksprache mit dem jeweils zuständigen Fachkoordinator notwendig machte. Die aus dem Korrekturbedarf, den Einschätzungen der Fachkoordinatoren zum Erhebungsaufwand und der Nützlichkeit der Daten für die Gutachter sich ergebenden Empfehlungen zur Verbesserung der Datengrundlage sind Gegenstand der Abschlussberichte beider Bewertungsgruppen.

Deutliche Unterschiede zwischen beiden Fächern der Pilotstudie bestehen hinsichtlich der verfügbaren Datengrundlage für bibliometrische Analysen. Die für die Chemie verwendeten Datenbanken aus dem Web of Science von Thomson Scientific decken die Zeitschriftenliteratur dieses Faches sehr gut ab und ermöglichen eine Zitationsanalyse. Es gab keine Einsprüche aus dem Fach gegen die Verwendung dieser Datenbanken. Da andere Publikationstypen als Zeitschriftenbeiträge in diesen Datenbanken nicht erfasst sind, wurde in der Datenerhebung ergänzend die Möglichkeit angeboten, andere Publikationen selbst einzugeben. Diese Ergänzungen machten mengenmäßig etwa 5 % der im Web of Science erfassten Zeitschriftenbeiträge aus. Überwiegend handelte es sich dabei um Beiträge zu Handbüchern, Lexika und Tagungsbänden.

Demgegenüber war die Abdeckung der Publikationen der teilnehmenden Soziologen durch die verwendeten Datenbanken SOLIS und CSA, die keine Zitationsanalysen unterstützen, lückenhaft. Der Umfang der ursprünglich in den Datenbanken recherchierten Publikationslisten wurde in der Korrekturrunde etwa verdoppelt. Da sich nicht alle Wissenschaftler an der Korrektur beteiligten, dürfte das Ausmaß der Unvollständigkeit der vorhandenen Datenbanken sogar noch höher liegen. Etwa ein Viertel der Nachmeldungen waren internationale Publikationen; drei Viertel hingegen Publikationen in Deutschland, die von der deutschen sozialwissenschaftlichen Literaturdatenbank SOLIS nicht erfasst werden. Ein wesentlicher Grund für die Unvollständigkeit der Datenbanken ist die hoch diversifizierte Publikationspraxis der Soziologen: Anders als in der Chemie sind Beiträge zu Sammelwerken mit ca. 45 % aller erfassten Publikationen der häufigste Publikationstyp, gefolgt von Zeitschriftenaufsätzen (34 %, davon ca. ein Viertel in internationalen Zitationsdatenbanken erfasst) und Monographien (7 %). Hinzu kommt, dass die Soziologie disziplinär weniger deutlich abgegrenzt ist. So, wie sich die Mehrheit der Forschungseinheiten als interdisziplinär bezeichnet, erfolgt auch ein großer Teil der Publikationen von Soziologen nicht in eindeutig als soziologisch zu identifizierenden Zeitschriften, sondern in einem breiten Spektrum von über 1.000 Zeitschriften unterschiedlichster Spezialisierung[14] und in Sammelwerken. In erheblichem Umfang erscheinen die Publikationen in Verlagen, deren Qualitätssicherungssystem für wissenschaftliche Publikationen unklar ist. In einem Fach mit einer solchen Publikationspraxis können die Publikationen nicht auf

[14] Während die für die Chemie analysierten ca.40.000 Zeitschriftenartikel sich auf gut 1.700 Zeitschriften verteilten, verteilten sich die nur ca. 4.000 soziologischen Artikel auf über 1.000 Zeitschriften. Um 50 % der Zeitschriftenartikel zu erfassen, müssten für die Chemie 47, für die Soziologie aber 76 Zeitschriften ausgewertet werden.

Basis weniger Datenbanken vollständig erfasst werden. Eine Vollerhebung setzt deshalb in diesen Fächern voraus, dass die betroffenen Wissenschaftler bzw. Forschungseinrichtungen an der notwendigen Ergänzung des Datenbestands, die durch eine zentrale Serviceeinrichtung organisiert werden muss, mitwirken.

### II.4. Erfahrungen aus dem Bewertungsvorgang

Die Unterteilung der Bewertungsphase in eine arbeitsteilige Vorbereitung durch zwei Berichterstatter je Einheit und eine plenare Diskussion aller Noten mit abschließender Konsistenzprüfung hat sich sehr bewährt. Das Gros der Forschungseinheiten konnte unter Berücksichtigung ihrer Spezialisierung und von Befangenheiten mit Berichterstattern aus der jeweiligen Bewertungsgruppe abgedeckt werden. Forschungseinheiten, denen ein Mitglied der Bewertungsgruppe angehörte, wurden in der Soziologie generell von einem externen Sondergutachter und dem Vorsitzenden der Bewertungsgruppe, in der Chemie in der Regel von einem externen Sondergutachter und einem Mitglied der Bewertungsgruppe bewertet. Einige sehr breit aufgestellte Forschungseinheiten wurden von drei Berichterstattern bewertet. In knapp 8 % der Fälle wurden Sondergutachter hinzugezogen, um Befangenheiten auszuschließen, hoch spezialisierte Forschungseinheiten zu bewerten oder zusätzliche Meinungen zu Fällen einzuholen, in denen ein Dissens zwischen den Berichterstattern von diesen nicht aufgehoben werden konnte. Vereinzelt wurden Sondergutachter auch in sehr großen Teilgebieten eingesetzt, um eine Überlastung der zuständigen Mitglieder der Bewertungsgruppe zu vermeiden.

Während die individuelle Bewertungsphase in der Chemie etwa sechs Wochen umfasste, dauerte sie in der Soziologie zwölf Wochen. Der größere Zeitbedarf in der Soziologie war wegen der Notwendigkeit, in großem Umfang exemplarische Publikationen der zu bewertenden Forschungseinheiten zu lesen, von vorneherein einkalkuliert worden, erwies sich aber auch deshalb als notwendig, weil die geringere Qualität der Publikationsdaten Nachrecherchen in größerem Umfang erforderlich machte.

Um die plenaren Abstimmungen zu erleichtern, wurde den Gutachtern vorab mitgeteilt, in welchen Fällen ihr Notenvorschlag von dem ihres jeweiligen Co-Berichterstatters abwich. Insgesamt bestand eine hohe Übereinstimmung zwischen den unabhängigen Urteilen der Gutachter – je nach Kriterium machten beide Berichterstatter in 75 – 86 % (Chemie) bzw. 71 – 89 % (Soziologie) der Fälle übereinstimmende

Vorschläge. Die Endergebnisse der Bewertung hängen nicht davon ab, welchen Berichterstattern eine Einheit zugeordnet wurde. Ein etwaiger Gutachterbias, der bei einer Einzelbegutachtung durchschlagen könnte, oder eventuelle Vorfestlegungen wurden durch detaillierte Erörterungen in den heterogen zusammengesetzten Gutachtergruppen neutralisiert und möglichen Konformitätstendenzen mittels kritischer Diskussionen sowie dadurch, dass wechselnde Gutachter in den plenaren Sitzungen die Rolle eines „advocatus diaboli“ übernahmen, vorgebeugt. Wichtig schien zudem, durch die Möglichkeit, einzelne Einheiten als „nicht bewertbar“ zu bezeichnen, einen Entscheidungszwang zu vermeiden.[15]

Bei einem empirisch fundierten Verfahren wie dem Forschungsrating können Fehler nicht vollständig ausgeschlossen werden. Auf Basis der vorhandenen Daten, der Gutachtervoten und deren Güteprüfung im Plenum kommen die Bewertungsgruppen zu dem Schluss, dass die Wahrscheinlichkeit von Fehlbewertungen relativ gering ist und deren Ausmaß, sollten sie entstanden sein, nicht über eine Notenstufe hinausgeht. Angesichts der Tatsache, dass bei unsicherer Datenlage oder unauflösbaren Dissensen die betreffenden Einheiten als „nicht bewertbar“ bezeichnet wurden, kann es sich jedoch nur um Einzelfälle handeln.

In der Chemie entschieden die Gutachter, die Forschungsqualität, für die die aussagekräftigsten Daten vorlagen, noch etwas differenzierter zu bewerten, indem zwischen „sehr gut“ und „exzellent“ eine weitere Skalenstufe „sehr gut bis exzellent“ eingeführt wurde. Schwierig erschienen ihnen demgegenüber die Kriterien in der Dimension Wissenstransfer, insbesondere das Kriterium „Wissensvermittlung und –verbreitung“. Wegen der vorwiegend qualitativen, zudem sehr heterogenen Datengrundlage entschied sich die Bewertungsgruppe Chemie, für das letztgenannte Kriterium nur eine Bewertung in den drei Stufen „unterdurchschnittlich, durchschnittlich, überdurchschnittlich“ vorzunehmen. Die Bewertungsgruppe Soziologie schloss sich dem an und verwendete die vereinfachte Skala auch für das Kriterium „Transfer in andere gesellschaftliche Bereiche“. Diese Änderung im laufenden Verfahren machte es erforderlich, einen Teil der Bewertungen zu wiederholen.

[15] Zu möglichen Fehlentscheidungen unter Entscheidungszwang und Stress vgl. Janis, Irving. L. (1982). Groupthink. Psychological studies of policy decisions and fiascoes. Boston: Houghton Mifflin.

Während der plenaren Beratungen zeigte sich, dass die Bewertung in einigen Sonderfällen[16] nicht möglich war. Um das Votum „nicht bewertbar“ in solchen Fällen kurz zu begründen, aber auch, um in Einzelfällen die Gründe für eine bestimmte Bewertung nachvollziehbar zu machen oder die Bewertung zu qualifizieren, bedienten sich beide Bewertungsgruppen des Instruments einer Kommentierung einzelner Bewertungsergebnisse.

### II.5. Zusammenfassung und Rezeption der Ergebnisse

Die Ergebnisse der Bewertung wurden durch die Steuerungsgruppe verabschiedet und veröffentlicht.[17] Die beiden Publikationen enthielten dabei jeweils die Bewertungen aller in dem jeweiligen Fach teilnehmenden Einrichtungen nach den sechs Kriterien des Forschungsratings sowie ein anonymisiertes Profil, das die Verteilung der gewichteten Forschungseinheiten der jeweiligen Einrichtung über die Notenstufen nach dem Kriterium „Forschungsqualität“ zeigt. Wie welche Forschungseinheiten abgeschnitten hatten, wurde aus Datenschutzgründen nicht veröffentlicht, sondern nur den jeweiligen Einrichtungen und ihren Landesministerien respektive ihren Trägereinrichtungen mitgeteilt.

In beiden Fächern der Pilotstudie und bei allen Kriterien wurde die Notenskala voll ausgeschöpft. In der Chemie streuen die Bewertungen bei allen Kriterien symmetrisch um die Note „gut“, die mit 25 – 40 % aller Bewertungen (je nach Kriterium) auch die häufigste Note ist. Die Spitzennote „exzellent“ wurde nach dem zentralen Kriterium „Forschungsqualität“ für 5 % aller Einheiten[18], nach den anderen, auf Einrichtungsebene bewerteten Kriterien in 6 – 12 % aller Fälle vergeben. In der Soziologie ist die Forschungsqualität im Mittel knapp mit „gut“ bewertet worden, hier haben 4 % der Forschungseinheiten „exzellent“ abgeschnitten. Auf Ebene der Einrichtungen sind je nach Kriterium 7 – 9 % der Fälle „exzellent“ bewertet worden. Die Nachwuchsförderung wird in der Soziologie im Mittel deutlich kritischer beurteilt als in der Chemie, wo sie zu einer der Stärken der deutschen Einrichtungen gezählt wird.

Die verschiedenen Kriterien, insbesondere die in der Dimension Forschung, hängen zum Teil inhaltlich miteinander zusammen, was jedoch nicht bedeutet, dass eines der Kriterien redundant wäre. Einrichtungen, die nach dem Kriterium Impact/Effekti-

---

16 Je nach Kriterium 2 – 6 % der Fälle in der Chemie, 4 – 7 % der Fälle in der Soziologie.
17 s. o., Fn. 8, S. 12.
18 Zusätzlich 7 % „sehr gut bis exzellent“.

vität im Mittelfeld liegen, können hinsichtlich der Forschungsqualität, der Effizienz, der Nachwuchsförderung oder des Wissenstransfers hervorragend abschneiden, wie auch umgekehrt die in der Forschung sehr sichtbaren Einrichtungen nicht immer die effizientesten oder diejenigen mit besonderen Leistungen in Nachwuchsförderung oder Transfer sind. Die große Bedeutung der Mehrdimensionalität für die Akzeptanz des Forschungsratings tritt in den Reaktionen der Einrichtungen auf die Veröffentlichung der Ergebnisse zutage, die sich häufig bewusst auf ein bestimmtes, für das Profil der jeweiligen Einrichtung zentrales Kriterium fokussieren.

Die Bewertung der Forschungsqualität der Forschungseinheiten fällt auch innerhalb der einzelnen Einrichtungen sehr differenziert aus; in mehr als der Hälfte aller Fälle – sowohl in der Chemie als auch in der Soziologie – erstrecken sich die Bewertungen der Forschungseinheiten einer Einrichtung auf mindestens drei Notenstufen. Im Verlauf der Pilotstudie Chemie wurde deutlich, dass ein großes Interesse daran bestehen könnte, nicht nur die Ergebnisse für ganze Einrichtungen, sondern auch die Bewertungen der einzelnen Forschungseinheiten zu veröffentlichen, die den Einrichtungen im Rahmen der Pilotstudie zum internen Gebrauch mitgeteilt, aber nur anonymisiert veröffentlicht wurden. Eine Veröffentlichung auf Ebene der Forschungseinheiten würde es den Einrichtungen erleichtern, die Ergebnisse im Vergleich mit den jeweils unmittelbaren Konkurrenten der einzelnen Forschungseinheiten zu interpretieren. Zudem wäre diese Information für externe Nutzer, bspw. für potentielle Kooperationspartner oder für Nachwuchswissenschaftler, von großem Nutzen. Gegen die Veröffentlichung der Ergebnisse von Forschungseinheiten spricht, dass diese Einheiten zum Teil aus weniger als drei (leitenden) Wissenschaftlern bestehen, so dass die Bewertungen potentiell personenbeziehbar sind und somit dem Datenschutz unterliegen. Datenschutzrechtlichen Bedenken könnte vorgebeugt werden, indem die Personenbeziehbarkeit vermieden oder vorab eine Zustimmung der Betroffenen zur Veröffentlichung eingeholt würde.

Ausgehend von den vorliegenden Ergebnissen kann die Frage gestellt werden, ob ein optimiertes Ranking, das quantitative Indikatoren mit Gewichtungsfaktoren zu einem Gesamtwert verrechnet, in der Lage wäre, das aufwendige Begutachtungsverfahren zu ersetzen. Dies ist für die Kriterien I bis IV[19] grundsätzlich denkbar, wobei

[19] Forschungsqualität, Impact/Effektivität, Effizienz, Nachwuchsförderung. Für die Kriterien Transfer in andere gesellschaftliche Bereiche sowie Wissensvermittlung und –verbreitung ist die Frage aufgrund der fehlenden quantitativen Daten nicht sinnvoll.

nach den besten bislang erprobten Modellen je nach Kriterium zwischen 20 und 30 %, bei der Forschungsqualität sogar 36 % der Fälle durch solch ein quantitatives Ranking um mindestens eine Skalenstufe anders bewertet werden als durch die Gutachter. Im Zuge einer weiteren Analyse des Nutzens eines Forschungsratings sollte diese Frage weiter verfolgt werden, auch wenn eine quantitative Bewertung prinzipielle Nachteile gegenüber einem „informed peer review“ selbst dann hat, wenn sie die Ergebnisse gut prognostiziert (vgl. B.I, S. 27; zur Vergleichbarkeit der Ergebnisse mit bereits publizierten Rankings vgl. Anhang, S. 47 ff.).

Über die Ergebnisse der Pilotstudie Chemie und über das neue Verfahren wurde in der Presse bereits umfassend berichtet. Die Mehrzahl der Berichte wies auf die Besonderheiten des Forschungsratings gegenüber herkömmlichen Rankings – insbesondere das Prinzip des „Informed Peer Review“ und die Mehrdimensionalität – hin; mehrere Kommentatoren begrüßten, dass der Wissenschaftsrat mit seinem Verfahren neue Standards für Leistungsvergleiche in der Wissenschaft aufstelle. Versuche, die Ergebnisse in schlagzeilenträchtige Hitlisten umzurechnen, sind bislang nicht bekannt. Schwierigkeiten bereitete allerdings die Tatsache, dass die Bewertungen nach den Kriterien der Transferdimension zum Teil auf unterschiedlichen Skalen erfolgten. Einige Medienberichte beschränkten sich deshalb auf die Bewertungen der Dimension Forschung.

Neben den veröffentlichten und den vertraulichen Bewertungsergebnissen erhielten die bewerteten Einrichtungen zum Abschluss der Pilotstudie auch die Datenberichte mit den statistischen Auswertungen der von ihnen eingelieferten Daten sowie ein Übersichtspapier, dass Angaben zum Erhebungsmodus und zu den nationalen Verteilungen der quantitativen Daten enthielt und den Gutachtern für ihre Bewertungen ebenfalls vorgelegen hatte. Detaillierte Nachfragen der Einrichtungen zu diesen Unterlagen machen deutlich, dass die Datenerhebung und die Analyse und Zusammenfassung der Daten durch die Geschäftsstelle für die teilnehmenden Einrichtungen zusätzlichen Nutzen generiert.

### II.6. Aufwand

Der Aufwand der Pilotstudie setzt sich aus drei Komponenten zusammen, die mit unterschiedlicher Genauigkeit beziffert werden können:

- Die direkten Kosten, die durch die Administration des Verfahrens einschließlich Sitzungsorganisation, Reisekosten und Aufwandsentschädigungen der Gutachter sowie durch die Publikations- und Zitationsanalysen einschließlich Lizenzgebühren entstanden. Sie betrugen ca. 1,1 Mio. Euro für Chemie[20] und Soziologie zusammen über die gesamte Laufzeit der Pilotstudie.
- Der Aufwand an Arbeitszeit der 15 bzw. 16 Gutachter je Fach, die nach eigenen Schätzungen im Laufe der Pilotstudie 4 – 5 (Chemie) respektive 8 – 10 Arbeitswochen (Soziologie) pro Person investiert haben. Der Unterschied zwischen beiden Fächern war neben der Tatsache, dass für die Soziologie eine umfassendere Lektüre von ausgewählten Publikationen notwendig war, vor allem darin begründet, dass die geringere Datenqualität in der Soziologie eine intensivere Auseinandersetzung mit den Rohdaten erzwang.
- Die bei den teilnehmenden Einrichtungen durch die Datenerhebung und die Korrektur der Publikationslisten entstehenden indirekten Kosten. Diese können nicht genau beziffert werden, da die Einrichtungen hierzu sehr unterschiedliche Angaben machen; nach pessimistischer Schätzung könnte der Aufwand pro Einrichtung bis zu zwei Personenmonate betragen haben, wobei große Unterschiede zwischen den einzelnen Einrichtungen berücksichtigt werden müssen und unklar ist, wie hoch der Anteil von Wissenschaftlern gegenüber Verwaltungsangestellten ist.

Diese Schätzungen beziehen sich auf die erstmalige Durchführung des Forschungsratings in Deutschland überhaupt und in den beiden genannten Fächern. Zu Faktoren, die den Aufwand des Forschungsratings bei einer Ausweitung oder Wiederholung des Verfahrens beeinflussen, vgl. B.II.6, S. 44 ff.

[20] Der Fonds der Chemischen Industrie hat mit ca. 25 % der anteiligen Kosten zur Pilotstudie Chemie beigetragen.

# B. Empfehlungen

## B.I. Empfehlungen zur Zukunft des Forschungsratings

Das Forschungsrating hat eine Reihe von Alleinstellungsmerkmalen, die es gegenüber den verbreiteten Rankings auszeichnen:

- Die Forschungsqualität wird durch „Informed Peer Review“ auf Basis quantitativer und qualitativer Vergleichsdaten differenziert und unter Berücksichtigung von Kontextinformationen vergleichend bewertet;
- die Fachgemeinschaften wirken an der Definition und Operationalisierung der Bewertungskriterien mit;
- durch den Ausweis der Differenzierung der Forschungsqualität innerhalb der einzelnen Einrichtungen haben die Ergebnisse einen höheren Informationsgehalt;
- durch die Bewertung nach mehreren Kriterien spiegeln sich unterschiedliche Leistungsprofile der Einrichtungen in den Ergebnissen wider;
- durch die Einbeziehung der außeruniversitären Forschungseinrichtungen, die in Fächern wie der Chemie eine große Rolle spielen, wird ein umfassendes Bild der deutschen Forschungslandschaft gezeichnet.

Die Pilotstudie hat anhand einer Natur- und einer Sozialwissenschaft gezeigt, dass das Verfahren für ein Forschungsrating durchführbar ist und zu aussagekräftigen Ergebnissen führt. Die einzelnen Kriterien sind dabei, wie in den Berichten der Bewertungsgruppen ausgeführt, unterschiedlich belastbar. Die Pilotstudie ist aus Sicht der Steuerungsgruppe Teil eines Lernprozesses, innerhalb dessen das Verfahren für ein Forschungsrating schrittweise weiterzuentwickeln ist. Im Zuge der Weiterentwicklung sollte im Dialog mit den Adressaten weiter geklärt werden, wie das Verfahren den bestmöglichen Nutzen entfalten und zugleich der Aufwand begrenzt werden kann.

Aufgrund seiner Alleinstellungsmerkmale kann das vom Wissenschaftsrat entwickelte Forschungsrating aus Sicht der Steuerungsgruppe im deutschen Wissenschaftssystem eine Reihe von Funktionen wahrnehmen, die durch bestehende Verfahren nicht zufriedenstellend erfüllt werden können:

- Das Forschungsrating gibt den wissenschaftlichen Einrichtungen Anstöße dazu, ihre eigene Strategie fortzuentwickeln, und kann der Erfolgskontrolle dienen.[21] Es hilft ihnen, sich in einem zunehmend wettbewerblich organisierten System ein erfolgversprechendes Profil zu geben. Dazu benötigen ihre Leitungsorgane verlässliche Bewertungen ihrer Stärken und Schwächen im Vergleich mit ihren unmittelbaren Wettbewerbern.[22]
- Das Forschungsrating liefert deutlich validere Informationen über die Qualität erbrachter wissenschaftlicher Leistungen als es bspw. die Auswertung von Drittmittelstatistiken und anderen quantitativen Indikatoren in herkömmlichen Rankings kann. Damit erfüllt es einen Bedarf der Wissenschaftspolitik an verlässlichen Informationen über die Leistungen wissenschaftlicher Einrichtungen, der in dem Maße steigt, in dem Mittel für die Wissenschaft stärker im Wettbewerb vergeben werden sollen.
- Indem das Forschungsrating Leistungen deutscher wissenschaftlicher Einrichtungen transparent und in einem Format darstellt, das mit international etablierten Bewertungsverfahren vergleichbar ist, erhöht es die internationale Sichtbarkeit der deutschen Wissenschaft. Damit bildet es auch ein Gegengewicht zu internationalen Rankings, die die außeruniversitäre Forschungslandschaft in Deutschland nicht berücksichtigen.
- Durch die fachspezifische Bewertung, die Differenzierung verschiedener Kriterien einschließlich einer auf den Personaleinsatz relativierten Effizienzbewertung und besonders durch die Bewertung einzelner Forschungseinheiten kann das Forschungsrating gute Forschungsleistungen auch außerhalb der national führenden Einrichtungen verlässlich identifizieren. Angesichts der derzeitigen Fokussierung auf internationale Exzellenz ist es wichtig, vielversprechenden Forschungsaktivitäten auch an kleineren Standorten mehr Aufmerksamkeit zu verschaffen, damit die wissenschaftliche Qualität in der Breite als Basis für Spitzenleistungen nicht verloren geht.
- Indem sie das Forschungsrating unterstützt und zum Anlass für eine kritische Selbstreflexion nimmt, trägt die Wissenschaft zu mehr Transparenz und zu einem

---

[21] Das Forschungsrating liefert keine Erklärungen dafür, warum eine Einrichtung mehr oder weniger leistungsfähig ist, und kann auch keine ex ante-Bewertung neuer Strategien leisten. Es ersetzt deshalb weder detaillierte Einzelevaluationen noch eine Beratung durch Beiräte.

[22] Für Vergleichszwecke ist es wichtig, möglichst detaillierte Informationen über die Bewertung von Wettbewerbern zu erhalten. Dies spricht für eine Veröffentlichung der Ergebnisse auch der einzelnen Forschungseinheiten, sofern die damit verbundenen datenschutzrechtlichen Bedenken ausgeräumt werden können (s. B.II.5, S.42 f.).

effizienten Mitteleinsatz bei. Das Verfahren hilft damit, eine aus Gesellschaft und Politik an die Wissenschaft herangetragene Rechenschaftspflicht zu erfüllen.

- Das Forschungsrating hebt die Bedeutung der Förderung des wissenschaftlichen Nachwuchses als eigenständige Aufgabe der wissenschaftlichen Einrichtungen hervor. Nachwuchswissenschaftler und fortgeschrittene Studenten aus dem In- und Ausland können mit Hilfe des Forschungsratings beurteilen, an welchen Einrichtungen ein passendes Umfeld für die ersten Schritte einer wissenschaftlichen Karriere besteht.
- In Fächern wie der Chemie erfüllt das Forschungsrating eine wichtige Orientierungsfunktion für potentielle Kooperationspartner aus der Industrie, deren Standortentscheidungen im Zuge der Globalisierung zunehmend auch von der Attraktivität des Forschungsumfeldes abhängig sind.

Gegenüber den bereits existierenden Rankings hat das Forschungsrating den Vorzug, Qualitätsurteile zu ergeben, die differenzierter und verlässlicher sind als auf rein quantitativer Datenbasis errechnete Ranglisten. Eine Analyse der Ergebnisse zeigt, dass gerade das zentrale Kriterium Forschungsqualität in mehr als einem Drittel der Fälle anders bewertet werden würde, wenn man die Bewertung durch eine bloße Gewichtung der quantitativen Indikatoren ohne Begutachtung zu berechnen versuchte. Auch Abweichungsraten von 20 – 30 %, wie sie für die übrigen Kriterien ermittelt wurden, sind inakzeptabel und zeigen, dass die Mitwirkung der Gutachter im „Informed Peer Review“ unverzichtbar ist. Andererseits wären ohne die umfassende Datengrundlage, die für das Forschungsrating erhoben wurde, auch erfahrene Gutachter nicht in der Lage, außerhalb der Gruppe der bekannten, besonders forschungsaktiven Einheiten verlässlich zu differenzieren. In einzelnen Fällen hat die Einschätzung der Daten die Gutachter auch dazu gebracht, bekannte Einheiten kritischer zu bewerten als aufgrund ihrer Reputation im Vorfeld zu erwarten war. Ein Mehrwert des Forschungsratings liegt also gerade in dem Zusammenwirken von Expertise und Erfahrung der Gutachtergruppe mit einer umfassenden, partiell statistisch auswertbaren, aber auch qualitative Komponenten umfassenden Datengrundlage.

Die Tatsache, dass das Forschungsrating ein Begutachtungsverfahren ist, darf nicht dahingehend missverstanden werden, es handele sich um eine Addition schematisierter Evaluationen der einzelnen Einrichtungen oder könne Evaluationen ersetzen. Der Wert des Verfahrens liegt in dem Vergleich zahlreicher Einrichtungen nach ein-

heitlichen Maßstäben. Es ist nicht seine Aufgabe, die Ursachen für die Leistungsfähigkeit einer bestimmten Einheit zu analysieren und Empfehlungen zu ihrer Verbesserung abzugeben.

Bei einer Verstetigung des Forschungsratings wäre zu beachten, dass Verfahren, die eine Stärkung des Leistungswettbewerbs im Wissenschaftssystem zum Ziel haben, neben den erwünschten auch nichtintendierte Anreizeffekte haben. Diese Effekte sollten anhaltend beobachtet und kritisch diskutiert werden. Selbst wenn es keine direkte Umrechnung von Indikatoren in Rangplätze gibt, werden bestimmte Daten allein durch die bloße Tatsache ihrer Erhebung so ausgezeichnet, dass es für die bewerteten Einrichtungen rational erscheint, Aufmerksamkeit und Ressourcen in die Optimierung der dadurch gemessenen Prozesse zu investieren. Der im Forschungsrating verfolgte Ansatz, mehrere Leistungsdimensionen ungewichtet nebeneinander zu bewerten und zudem der Bewertung neben quantitativen auch qualitative, auf offenen Fragen beruhende Informationen zugrunde zu legen, ist ein Weg, derartige homogenisierende Effekte nach Möglichkeit zu begrenzen. Bei einer zu weitgehenden Vereinfachung ginge dieser auch für die Akzeptanz in der Wissenschaft entscheidende Vorzug des Verfahrens verloren.

Die vorliegenden Erfahrungen sprechen dafür, dass das Forschungsrating mit entsprechenden Verfahrensanpassungen im vorgesehenen Rahmen auch auf andere Wissenschaftsbereiche übertragen werden kann. Die größten Modifikationen sind dabei zu erwarten, wenn Fachgebiete aus den Geistes- und den Technikwissenschaften bewertet werden, da diese sich sowohl hinsichtlich der fachinternen Publikations- und Kommunikationswege als auch hinsichtlich ihrer für die Transferdimension relevanten Beziehungen zu anderen gesellschaftlichen Bereichen am stärksten von den Natur- und den Sozialwissenschaften unterscheiden. Dies legt nahe, solche Fachgebiete bei einer Fortentwicklung des Verfahrens als nächste in Angriff zu nehmen.

Im Rahmen der Fortentwicklung des Forschungsratings sollte der Dialog mit den Adressaten gesucht werden, um besser einschätzen zu können, wie die veröffentlichten Ergebnisse, auch im Vergleich mit eventuell vorhandenen internen Evaluationen, tatsächlich genutzt werden. Im Anschluss daran sollte die Empfehlung der Bewer-

tungsgruppe Chemie, die Chemie in fünfjährigem Abstand erneut zu bewerten und parallel auch die Biologie und die Physik zu bewerten, erneut geprüft werden.

Ein wesentlicher Vorzug des in der Pilotstudie durchgeführten Forschungsratings ist, dass es von der Wissenschaft mit gestaltet wird. Seine Akzeptabilität wird mittelfristig davon abhängen, dass auch unerwünschte Folgen, die sich aus den Anreizeffekten solcher Bewertungen ergeben können, bei der Weiterentwicklung des Verfahrens berücksichtigt werden. Die Steuerungsgruppe empfiehlt dem Wissenschaftsrat deshalb, die Untersuchung der Folgen des Forschungsratings und verwandter Bewertungsverfahren zu unterstützen.

## B.II. Zur Optimierung des Verfahrens für ein Forschungsrating

### II.1. Zu Organisation und Ablauf

Größe und Zusammensetzung der Bewertungsgruppen in der Pilotstudie haben sich bewährt. Für kleinere Fächer sind ggf. auch etwas kleinere Gruppen denkbar; größere Gruppen werden nicht empfohlen. Die internationale Erfahrung der Gutachter ist gerade in den Fächern, für die internationale Vergleichsdaten nicht vorliegen, besonders wichtig, um eine Bewertung nach internationalen Maßstäben sicherzustellen. Der Anteil von Gutachterinnen sollte gesteigert werden. Der Rückgriff auf Vorschläge der Wissenschaftsorganisationen und die Einbeziehung der Fachgesellschaften in die Gutachterrekrutierung haben neben dem pragmatischen Nutzen auch zur Akzeptanz des Verfahrens in den Fächern wesentlich beigetragen und sollten unbedingt beibehalten werden. Das Renommee der Gutachter innerhalb ihres Faches ist für die Akzeptanz des Verfahrens generell von großer Bedeutung.

Die umfassende Unterstützung durch die Geschäftsstelle war nach Aussagen der Bewertungsgruppen von großer Bedeutung für das Gelingen des Forschungsratings und muss im Falle einer Fortsetzung auch künftig gewährleistet sein, um den Arbeitsaufwand für die einzelnen Gutachter in akzeptablen Grenzen zu halten. Dies ist bei einer Wiederholung im gleichen Fach besonders wichtig, weil dann der Reiz des ersten Experiments fehlt. Die Bereitschaft von renommierten Fachwissenschaftlern, als Gutachter an einem Forschungsrating mitzuwirken, hängt entscheidend davon ab, dass ihnen eine gute Unterstützung und eine saubere Datengrundlage zugesichert werden kann. Die in der Pilotstudie geleistete und für Fächer, die zum ersten

Mal einer solchen Bewertung unterzogen werden, notwendige konzeptionelle Vorarbeit kann nur durch die Geschäftsstelle übernommen werden, wenn für jedes Fach ein wissenschaftlicher Mitarbeiter zuständig ist. Mit der wiederholten Bewertung eines Faches ist eine Routinisierung und eine Verringerung des personellen Aufwands erwartbar.

Der Ablauf der Pilotstudie hat sich im wesentlichen bewährt und sollte auch für künftige Forschungsratings beibehalten bleiben. Für jedes Fach, das erstmalig einer Bewertung dieser Art unterzogen wird, ist eine ausreichende Zeit für die Indikatorenentwicklung wichtig, damit die Möglichkeit besteht, neu entwickelte Indikatoren gegebenenfalls in einem Pretest zu erproben. Aus Sicht der bewerteten Einrichtungen war problematisch, dass die Aufforderung, Forschungseinheiten zu bilden, bereits an sie erging, bevor sie genau wussten, wie die Forschungsleistungen der Einheiten bewertet werden würden. Solche Vorgriffe sollten künftig vermieden werden. Kritisiert wurde zudem die zu geringe Vorlaufzeit für die Datenerhebung (s. II.3, S. 36 f.). Den Einrichtungen mehr Zeit für die Vorbereitung zu geben, macht es möglich, die eigentliche Datenerhebung in einem festen Zeitfenster unmittelbar nach dem Stichtag vorzunehmen und Nachfristen nur noch in Ausnahmefällen einzuräumen. Dies führt auch dazu, dass die Daten zum Zeitpunkt der Bewertung aktueller sind.

## II.2. Zum Gegenstand der Bewertung

### a) Fachgebiete und Interdisziplinarität

Für eine vergleichende Bewertung von Forschungsleistungen ist es unverzichtbar, geeignete Vergleichsgruppen zu definieren. Für die Pilotstudie wurden diese Vergleichsgruppen anhand von Fächern definiert. Dafür spricht, dass wissenschaftliche Qualitätsstandards in erster Linie in den Fachgemeinschaften bestimmt werden. Dagegen spricht, dass fachübergreifende Einrichtungen in Fächer aufgeteilt werden müssen und das Geschehen zwischen den Fächern nicht mitbewertet wird. Zuordnungsprobleme treten umso häufiger auf, je differenzierter die zugrunde gelegte Fachsystematik ist; andererseits erhöht eine zu grobe, nur wenige Bereiche unterscheidende Systematik die Gefahr einer verzerrten Bewertung, weil einheitliche Maßstäbe an sehr unterschiedliche Forschungspraktiken angelegt werden. Dieses Spannungsverhältnis gilt es im Auge zu behalten, wenn das Forschungsrating fortgesetzt werden soll.

Die in der Pilotstudie vereinzelt aufgetretenen Zuordnungsprobleme waren überwiegend auf eine unvollständige Informationsgrundlage zurückzuführen: Wissenschaftler waren unsicher, welche Alternativen zu einer Einordnung in die Chemie respektive die Soziologie bestanden. Diese Ungewissheiten ließen sich bei einer dauerhaften Etablierung des Verfahrens vermeiden, wenn vor Beginn der Erhebungen eine Taxonomie aller in Frage kommenden Fachgebiete publiziert würde.[23] Ausgangsbasis einer solchen Taxonomie sollte die Systematik der aktuell 48 Fachkollegien der DFG mit den ihnen zugeordneten Fächern sein. Dabei ist darauf zu achten, dass die Fachgebiete nicht zu kleinteilig definiert werden. Für Fachgebiete, die wesentlich kleiner sind als die Soziologie[24], steht der Aufwand eines Forschungsratings in keinem günstigen Verhältnis mehr zum Ertrag. In diesen Fällen sollten benachbarte Fächer nach Möglichkeit zu einem gemeinsamen Fachgebiet zusammengefasst werden; dies würde zugleich auch die Probleme der Zuordnung respektive der Interdisziplinarität reduzieren. Auf der anderen Seite setzt die Gutachterbelastung der Größe von Fachgebieten, die mangels Zitationsdaten nur durch eine Lektüre von ausgewählten Publikationen bewertet werden können, eine obere Grenze. Die vom Wissenschaftsrat 2004 avisierte Obergrenze von max. 50 Fachgebieten erscheint nach wie vor realistisch.

Bei der Definition der Fächer ist ferner auch die an den Universitäten typischerweise vorherrschende Abgrenzung der Organisationseinheiten zu berücksichtigen, da eine Datenerhebung quer zu den Organisationseinheiten für die Einrichtungen aufwendiger ist und die Ergebnisse weniger steuerungsrelevant sind. Um einen für alle Beteiligten akzeptablen Kompromiss zu finden, sollte der Entwurf einer Taxonomie in einem öffentlichen Konsultationsverfahren mit Universitäten, Wissenschaftsorganisationen und Fachgesellschaften erörtert werden, bevor er einem Forschungsrating mehrerer Fächer zugrunde gelegt wird. Ein geeigneter Zeitpunkt wäre die Vorbereitung einer möglichen Ausweitung des Verfahrens auf die Naturwissenschaften Biologie, Chemie und Physik.

Interdisziplinäre Forschungsaktivitäten sollten nach Möglichkeit im Zusammenhang bewertet und nicht künstlich auseinander dividiert werden. Um Schwierigkeiten der Bewertung vorzubeugen, sollten auch Wissenschaftler aus Randgebieten des jewei-

[23] vgl. Wissenschaftsrat 2004, S. 44.

[24] Die Soziologie hat in der DFG kein eigenes Fachkollegium, sondern ist eines der dem Fachkollegium „Sozialwissenschaften“ zugeordneten Fächer.

ligen Faches in die Bewertungsgruppen berufen werden. Die Bewertungsgruppen haben in der Pilotstudie gute Erfahrungen damit gemacht, für hoch spezialisierte – darunter auch interdisziplinäre – Forschungseinheiten Sondergutachter hinzuzuziehen.

**b) Forschungseinheiten**

Bei der Abgrenzung der Forschungseinheiten sind die beiden in der Pilotstudie untersuchten Fächer unterschiedlich vorgegangen. Die namentliche Erfassung der zugehörigen Wissenschaftler, die auch Ausgangspunkt der anschließenden Publikationsrecherche war, hat sich bewährt. Wegen der Verbindung zur Publikationsrecherche sollte die Erfassung der Personen an dieser Stelle jedoch nicht auf bestimmte Personalkategorien beschränkt werden, wie es in der Chemie (Professoren und Gruppenleiter) zunächst der Fall war, sondern die Meldung aller selbständig publizierenden Wissenschaftler möglich sein. Unterschiede gab es vor allem hinsichtlich der Bildung von Forschungseinheiten.

Angesichts der Kleinteiligkeit der in der Soziologie definierten Forschungseinheiten wurde in der Bewertungsgruppe Soziologie der Vorschlag diskutiert, künftig – ähnlich wie die Bewertungsgruppe Chemie – vorzugeben, dass nur Organisationseinheiten auf Institutsebene (Institut für Soziologie, Institut für Medizinische Soziologie etc.), nicht jedoch einzelne Professuren als Forschungseinheiten angemeldet werden können. Auf diese Weise würde sich die Zahl der Forschungseinheiten in der Soziologie um mehr als die Hälfte reduzieren lassen; zudem sind solche größeren Einheiten ein Gegenstand strategischer Steuerung, wie sie durch das Forschungsrating unterstützt werden soll, und können durch eine Bewertung in ihren Akteursqualitäten gestärkt werden. Die Bewertungsgruppe ist jedoch zu dem Schluss gekommen, dass die in der Pilotstudie erfasste, kleinteilige Struktur den gegenwärtigen Zustand des Fachs widerspiegelt. Die Entwicklung größerer, durch gemeinsame Forschungsprogramme, eine geteilte Infrastruktur und größere Kontinuität gekennzeichnete Einheiten in der Soziologie wäre wünschenswert, sie allein für die Zwecke des Forschungsratings als fiktive Einheiten zu definieren, würde jedoch aus Sicht der Bewertungsgruppe auch angesichts der oft erheblichen Leistungsunterschiede den Informationsgehalt der Ergebnisse zu sehr verringern.

Mit Blick auf die Akzeptanz in der Wissenschaft sind auch künftig Spielräume für unterschiedliche, den jeweiligen Kooperationspraxen angemessene Vorgaben für die Forschungseinheiten in den verschiedenen Wissenschaftsgebieten sinnvoll. Allerdings war die Varianz innerhalb der einzelnen Fächer aufgrund des experimentellen Charakters der Pilotstudie und der damit verbundenen Unsicherheit für die teilnehmenden Einrichtungen größer als unbedingt erforderlich. Eine bessere Standardisierung sollte dadurch sichergestellt werden, dass die endgültige Einteilung einer Einrichtung in Forschungseinheiten mit den Gutachtern frühzeitig abgestimmt wird.

### c) Außeruniversitäre Forschungseinrichtungen

Die Beteiligung außeruniversitärer Forschungseinrichtungen am Forschungsrating ist einer der großen Vorzüge dieses Verfahrens gegenüber den bestehenden nationalen und internationalen Rankings. In vielen Fachgebieten haben wie in der Chemie, aber anders als in der Soziologie außeruniversitäre Einrichtungen erheblichen Anteil an Umfang und Qualität der deutschen Forschungsleistungen. Die besten außeruniversitären Institute können zudem geradezu als Maßstab für internationale Forschungsqualität dienen und helfen somit, die Skala für die Bewertung zu kalibrieren. Ohne die außeruniversitären Institute würde das durch ein Forschungsrating gezeichnete Bild der deutschen Forschungslandschaft erheblich an Wert verlieren.

Außeruniversitäre Einrichtungen sind besonders häufig fachübergreifend organisiert und sehen einen besonderen Wert in dieser Organisation. Es muss weiter untersucht werden, wie dies im Rating berücksichtigt werden kann.

Spezifische Lösungen sind für Forschungseinheiten erforderlich, deren wissenschaftliche Leiter durch gemeinsame Berufung an einer Universität und einer außeruniversitären Einrichtung zugleich tätig sind. Das den acht am Pretest teilnehmenden Einrichtungen gemachte Angebot, „gemeinsame Einheiten“ zweier Partnereinrichtungen anmelden zu können, für die ein integrierter Datensatz produziert, nur einmal bewertet und dann beiden Einrichtungen zu einem von ihnen zu vereinbarenden Prozentsatz angerechnet werden sollte, wurde nur ein einziges Mal wahrgenommen. Andererseits hatten einige Einrichtungen offenkundig Schwierigkeiten mit der für die Haupterhebung gewählten Alternative, die jeweiligen wissenschaftlichen Leiter beiden Einrichtungen zuzurechnen, die Forschungsleistungen aber klar aufzutrennen. Größere, nicht im einzelnen nachvollziehbare Überlappungen der Daten führten in

Einzelfällen dazu, dass die betroffenen Forschungseinheiten nicht bewertbar waren. Im Rahmen der Weiterentwicklung des Forschungsratings sollten beide Alternativen – „gemeinsame Forschungseinheit“ oder „Auftrennung“ – angeboten und die Konsequenzen vorab klar kommuniziert werden. Auf Basis der Erfahrungen aus einem solchen Optionsmodell sollten mittelfristig Regeln dafür entwickelt werden, wann welche Option zu wählen ist.

### II.3. Zur Datenerhebung und -analyse

Eine Optimierung der Datenerhebung muss zwei Ziele verfolgen: die Datenqualität weiter zu verbessern, um die Bewertung durch die Gutachter zu erleichtern, und die Arbeitsbelastung für die Einrichtungen, die die Daten bereitstellen, zu reduzieren. Eine sorgfältige Vorbereitung der Datenerhebung für jedes Fach durch Bewertungsgruppe und Geschäftsstelle, die auch Pretests für neue Indikatoren umfassen können muss, wäre deshalb bei einer Fortsetzung des Forschungsratings unabdingbar. Mittelfristig sollte der Aufwand für die Einrichtungen auch dadurch begrenzt werden, dass sich Einrichtungen, die Forschungsdaten erheben, über die Erhebungsformate stärker abstimmen und so eine Mehrfachnutzung der Daten ermöglichen.

Die Indikatorenentwicklung mündet in der Aufstellung einer Bewertungsmatrix für jedes Fach, in der die einzelnen Kriterien durch sogenannte Bewertungsaspekte näher bestimmt und diesen Indikatoren zugeordnet werden. Dieses Dokument strukturiert die nachfolgende Datenerhebung und –analyse.

Die Datenerhebung in den einzelnen Einrichtungen sollte auch künftig in der Verantwortung eines von der Einrichtung zu benennenden Fachkoordinators liegen. Es hat sich bewährt, einen Fachwissenschaftler mit dieser Aufgabe zu betrauen, wobei es wichtig war, dass der Fachkoordinator von der Verwaltung seiner Einrichtung Unterstützung erhielt. Den Einrichtungen sollte davon abgeraten werden, die Datenerhebung an jüngere wissenschaftliche Mitarbeiter zu delegieren, die weder über die Forschungsschwerpunkte der Forschungseinheiten ihres Faches aus dem Erhebungszeitraum gut informiert sein können noch wissen, welche Datenquellen an ihrer Einrichtung existieren.

Für die Tätigkeit der Fachkoordinatoren ist eine rechtzeitige, gründliche und verlässliche Information über alle Details des Verfahrens unabdingbar. Die Datenerhebung

sollte deshalb erst beginnen, nachdem die Bewertungsmatrix veröffentlicht worden ist. Dabei ist zu beachten, dass aus Sicht der bewerteten Einrichtungen bereits die Erfassung der Forschungseinheiten Teil der Datenerhebung ist. Hilfreich wären Informationsveranstaltungen vor Beginn der Datenerhebung. Eine von vielen Fachkoordinatoren gewünschte längere Vorlaufzeit – mehrfach wurde vorgeschlagen, die Bewertungsmatrix und die darauf basierenden Fragebögen nach Vorbild der Research Assessment Exercise bereits vor Beginn des Erhebungszeitraums zu publizieren, um ein Sammeln der Daten im laufenden Prozess zu ermöglichen – würde dazu beitragen, dass die eigentliche Datenerfassung zügig erfolgen kann. Neben einer Verringerung des Aufwands hätte dies den Vorzug, dass die Daten zum Zeitpunkt der Bewertung durch die Gutachter aktueller sind als dies im Rahmen der Pilotstudie der Fall war. Bei Einführung eines rollierenden, nach und nach alle Fächergruppen erfassenden Systems sollte deshalb eine schrittweise Verlängerung der Vorlaufzeit für die Datenerhebung angestrebt werden. Zugleich ist in diesem Rahmen eine weitere Standardisierung der Erhebungsformate anzustreben, um eine Mehrfachverwendung erhobener Daten zu ermöglichen und die Einrichtungen in den Stand zu versetzen, eine Vorratsdatenhaltung für verschiedene Zwecke zu betreiben.

Publikationsdaten wurden in der Pilotstudie nicht direkt bei den Einrichtungen erhoben, sondern zunächst in vorhandenen Datenbanken recherchiert und dann gemeinsam mit den Fachkoordinatoren korrigiert. Dieses Vorgehen ist für Fächer mit einer internationalisierten Publikationspraxis wie der Chemie etabliert und führt zu guten, verlässlichen und hohe Akzeptanz findenden Ergebnissen. Demgegenüber war die Publikationsrecherche in vorhandenen Datenbanken für die Soziologie deutlich weniger erfolgreich. Neben dem geringeren Grad der Internationalisierung der Soziologie – 85 % der letztlich erfassten Publikationen sind im deutschen Sprachraum erschienen – und der unschärferen disziplinären Abgrenzung hat dies auch damit zu tun, dass Zeitschriftenartikel nur etwa ein Drittel der soziologischen Literatur ausmachen, Monographien und Sammelbände aber systematisch schwieriger zu erfassen sind. Dieser Befund dürfte sich auf viele andere geistes- und sozialwissenschaftliche Fächer mit einer überwiegend nationalen, hinsichtlich der Publikationsorgane hoch diversifizierten Publikationskultur übertragen lassen. Die Ausweitung des Forschungsratings auf weitere dieser Fächer wird deshalb die Notwendigkeit mit sich bringen, Literaturdaten unter Mitwirkung der betroffenen Wissenschaftler zu korrigieren und in erheblichem Umfang neu zu erfassen.

Aus Gründen des Aufwands ist in diesen Fächern eine selektive Vorgehensweise – z. B. eine Beschränkung auf die in bestimmten Datenbanken erfassten Zeitschriften – denkbar. Dabei ist aber dafür Sorge zu tragen, dass die dadurch bewirkten Anreizeffekte nicht zu einer unerwünschten Veränderung der Kommunikationsstruktur der betreffenden Wissenschaft und damit etwa zu einer Benachteiligung interdisziplinärer oder anwendungsorientierter Forschung führt. Vor diesem Hintergrund ist eine Vollerfassung der wissenschaftlichen Publikationen grundsätzlich wünschenswert. Der Zusammenhang mit der Leistungsbewertung stellt einen erheblichen Anreiz für Wissenschaftler eines Fachs dar, sich daran zu beteiligen. Dies kann zu einer deutlichen Verbesserung der Literaturdaten führen, die dem Fach zugute kommen sollte. Deshalb sollte nach Möglichkeit angestrebt werden, die Literaturdaten für solche Fächer in Kooperation mit einer Institution zu erheben, die die Daten nach Abschluss des Ratings weiter pflegt und für wissenschaftliche Nutzer verfügbar hält.

Die im Rahmen der Pilotstudie Chemie für die Zitationsanalyse verwendeten Indikatoren entsprechen dem internationalen Stand der bibliometrischen Forschung. Angesichts der beabsichtigten Wirkung des Forschungsratings verbietet es sich, experimentelle Indikatoren zugrunde zu legen, deren Verhalten noch nicht umfassend erprobt ist. Sollte sich jedoch zeigen lassen, dass neuartige Indikatoren valider und/oder manipulationsresistenter sind, sollte die bibliometrische Basis künftiger Bewertungen entsprechend erweitert werden. Es muss auch in einem etablierten Ratingverfahren Spielräume für die probeweise Bestimmung von Indikatoren geben, die bei der endgültigen Bewertung nicht verwendet werden. Ein besonderes Desiderat ist die Entwicklung von Indikatoren, die die Rezeption von Sammelbänden und Monographien widerspiegeln, da eine Beschränkung auf Artikel in Zeitschriften eine unangemessene Einflussnahme auf die Publikationskultur von Geistes- und Sozialwissenschaften bedeuten würde. Wie schon in der Pilotstudie sind auch bei künftigen Forschungsratings vor einer Auftragsvergabe Datenbanken konkurrierender Anbieter auf die Vollständigkeit, Qualität und Aussagekraft der Daten zu prüfen.

Der in der Pilotstudie geltende Erhebungszeitraum von fünf Jahren hat sich bewährt. Kürzere Zeiträume bringen die Gefahr mit sich, dass zufällige Schwankungen der Forschungsleistung auf die Bewertung durchschlagen, längerfristige Forschungsvorhaben systematisch benachteiligt sowie Aufwand und Dichte der Bewertungen zu groß werden.

In der Pilotstudie Chemie war es für manche Einrichtungen schwierig, Angaben über die Leistungen von Wissenschaftlern zu machen, die sie zum Stichtag bereits verlassen hatten. Auch in der Soziologie, in der nach dem „Current Potential“-Prinzip erhoben wurde, war die rückwirkende Datenerhebung teilweise problematisch. Der Grund war hier nicht, dass betroffene Wissenschaftler die Einrichtung verlassen hatten, sondern dass es keine institutionellen Prozeduren für die Datensammlung und –speicherung gab und, unter der „Current Potential“-Perspektive, auch nicht geben konnte, so dass häufig auf die Aufzeichnungen der einzelnen Wissenschaftler bzw. Arbeitsgruppen zurückgegriffen werden musste. Die Verwaltungen mehrerer Universitäten haben erklärt, künftig zumindest einen Grundbestand an Daten in dem für das Forschungsrating verwendeten Format kontinuierlich fortschreiben zu wollen. Andere Einrichtungen tun dies bereits jetzt im Rahmen ihrer regelmäßigen Forschungsberichterstattung. Ein solcher, von der Institution gepflegter Datenbestand korrespondiert der „Work Done At“-Perspektive und würde den Aufwand für künftige Datenerhebungen reduzieren. Da die Daten gleichzeitig auch für die Selbststeuerung der Einrichtungen und ihre Forschungsberichterstattung sowie für etwaige Evaluationen genutzt werden können, ist ein zusätzlicher Nutzen für die Einrichtungen zu erwarten. Während die Steuerungsziele durch eine „Current Potential“-Bewertung besser erfüllt werden, spricht der Aufwand deshalb eher für eine „Work Done At“-Bewertung, die im Rahmen des Peer Review auch aktuelle Veränderungen berücksichtigen kann.

Falls das Ratingverfahren auf weitere Fächer ausgedehnt wird, sollte die technische Realisierung der Erhebung verbessert werden, um die Dateneingabe seitens der Fachkoordinatoren zu erleichtern und so die Akzeptanz für das Verfahren zu steigern. Anzustreben ist die Entwicklung eines mit üblichen Dateiformaten kompatiblen Online-Systems mit Modulcharakter, das an die Bedürfnisse der verschiedenen Fächer leicht angepasst werden kann.

### II.4. Zu den Bewertungskriterien und zum Bewertungsvorgang

Das gegenüber dem ursprünglichen Vorschlag des Wissenschaftsrats vereinfachte Kriterienraster (vgl. Abb. 2, S. 14) hat sich in der Pilotstudie bewährt und sollte im wesentlichen beibehalten werden.

Für das Kriterium „Impact/Effektivität“, dessen Bedeutung für viele Nutzer schwer nachvollziehbar war, sollte ein anderer Begriff gewählt werden.

Hinsichtlich der Bewertung der Effizienz ist von verschiedenen Seiten empfohlen worden, auch Unterschiede in der Belastung durch Aufgaben neben der Forschung, beispielsweise durch die Lehre, Unterschiede der verfügbaren Forschungsinfrastruktur sowie des Ressourcenbedarfs unterschiedlicher Forschungsrichtungen innerhalb eines Fachs zu berücksichtigen. Im Rahmen der Weiterentwicklung des Verfahrens sollte geprüft werden, ob diese drei Faktoren mit vertretbarem Aufwand berücksichtigt werden können und ob Aussagen zur Effizienz ohne Berücksichtigung dieser Faktoren belastbar sind. Möglichkeiten, die Lehrbelastung empirisch zu gewichten, sind bereits im Pretest erprobt und als zu aufwendig verworfen worden. Belastbare, einrichtungsübergreifend vergleichbare Zahlen zu den tatsächlich mit der Erbringung bestimmter Forschungsleistungen verbundenen (Voll-)Kosten werden auf absehbare Zeit nicht zur Verfügung stehen.

Ein zentrales Desiderat für die Bewertung der Nachwuchsförderung sind Daten über den Verbleib der Promovierten. Bessere Daten darüber würden helfen, den Erfolg der Nachwuchsförderung auch qualitativ einzuschätzen, und würden zudem dazu führen, dass auch außerakademische Erfolge besser berücksichtigt werden. Die Hochschulen und außeruniversitären Forschungsinstitute sollten deshalb ermutigt werden, regelmäßige und systematische Absolventenstudien vorzunehmen.

Der Versuch der beiden Bewertungsgruppen, für die Kriterien der Dimension „Transfer“ auf eine gemeinsame Formulierung zu kommen, hat sich im Rückblick als unnötig erwiesen. In dieser Dimension sollten die Kriterien künftig stärker an die fachspezifischen Gepflogenheiten angepasst werden, wie der Wissenschaftsrat bereits in seinen Empfehlungen zu Rankings im Wissenschaftssystem vorgesehen hat.[25]

Die fünfstufige Bewertungsskala von „nicht befriedigend“ bis „exzellent“ hat sich ebenfalls bewährt. Die Reduzierung der Bewertungsskala auf drei Stufen in einem der (Chemie) bzw. beiden (Soziologie) Kriterien der Dimension Transfer war im Rahmen der Pilotstudie sinnvoll, dies hat jedoch die Nachvollziehbarkeit der Ergebnisse aus Sicht mancher Nutzer beeinträchtigt und dazu geführt, dass die Bewertungen in der Transferdimension weniger beachtet wurden. Bei der erstmaligen Anwendung des

[25] l.c., S. 47 Fn. 46.

Verfahrens in einem neuen Fach kann ein solcher Schritt zunächst unumgänglich sein; in der Regel sollte jedoch versucht werden, durch die Entwicklung entsprechend aussagekräftiger Indikatoren eine gleich differenzierte Bewertung aller Kriterien zu ermöglichen. Die Bewertungsgruppe Chemie hat in ihrem Abschlussbericht dazu Vorschläge gemacht. In der Transferdimension kann auch die Definition der Kriterien je nach den für das Fach typischen Verbindungen zu anderen gesellschaftlichen Praxen, die sich in der Verfügbarkeit von Indikatoren niederschlagen, angepasst werden. Die Bewertungsgruppe Soziologie schlägt in diesem Zusammenhang eine Zusammenfassung der beiden Transferkriterien vor, die ihrer Einschätzung nach auch für andere Sozial- und Geisteswissenschaften angebracht sein dürfte.

Die Pilotstudie hat bestätigt, dass Peer Review für eine verlässliche Bewertung der Qualität von Forschungsleistungen unverzichtbar ist. Dies ist unmittelbar evident in Fächern wie der Soziologie, für die valide Zitationsanalysen auf Basis der verfügbaren Datenbanken nicht flächendeckend möglich sind und in denen deshalb die Qualitätsbewertung durchgängig die Lektüre ausgewählter Publikationen voraussetzt. Auch in der Chemie beruht das Qualitätsurteil der Gutachter jedoch wesentlich auf qualitativen Informationen wie den Publikationslisten, den Selbstbeschreibungen der Einheiten und in vielen Fällen ebenfalls der Lektüre von Publikationen. Zudem wurden die Zitationsindikatoren von den Gutachtern kritisch überprüft und bisweilen unter Rückgriff auf die Rohdaten korrigiert. So ist es nicht unerwartet, dass der Versuch, die Bewertungsergebnisse durch eine Gewichtung der verwendeten quantitativen Daten statistisch zu prognostizieren, in 20 – 36 % der Fälle zu abweichenden Bewertungen führt. Ein rein indikatorenbasierter Leistungsvergleich kann das Forschungsrating nicht ersetzen.

Das Forschungsrating ist nur durchführbar, wenn die Bewertungsarbeit unter den Gutachtern aufgeteilt wird. Die arbeitsteilige Vorbereitung mit zwei Berichterstattern je Bewertung hat sich bewährt und liefert mit der Gutachterübereinstimmung zugleich ein Maß für die Reliabilität der Bewertung, das in der Pilotstudie, gemessen an den aus der Peer Review-Forschung bekannten Werten, hoch ausgefallen ist.[26] Die Be-

[26] S. a. A.II.4. Dazu Bornmann & Daniel (2003): „Begutachtung von Fachkollegen in der Wissenschaft“, in Schwarz & Teichler (Hg.) „Universität auf dem Prüfstand“, Campus Verlag, Frankfurt: S. 207 – 225; vgl. aber Hartmann & Neidhardt (1990): „Peer Review at the Deutsche Forschungsgemeinschaft“, Scientometrics 19, 419 – 425, wo bezüglich der Förderempfehlungen von ähnlich hohen Raten der Gutachterübereinstimmung berichtet wird wie in der Pilotstudie bezüglich der differenzierten Notenvorschläge erzielt wurden. Ein hoher Konsens kann auch Folge eines Konformitätsdrucks sein, vgl. dazu Janis (1982). Groupthink. Psychological studies of policy decisions and fiascoes. Boston: Houghton Mifflin. Zu den Maßnahmen, die dagegen getroffen wurden, vgl. A.II.4, S. 20 ff.

wertungsgruppen sollten auch künftig die Möglichkeit haben, Sondergutachter einzubeziehen, wenn ihnen dies zur Bewertung von hoch spezialisierten Forschungseinheiten oder zur Vermeidung von Befangenheiten notwendig erscheint. Die Regel, dass die Bewertungsvorschläge der Berichterstatter von der gesamten Bewertungsgruppe entscheiden werden müssen, sollte unbedingt beibehalten werden.

### II.5. Zur den Ergebnissen und ihrer Nutzung

Die Ergebnisse der Bewertungen beider Fächer sind sowohl hinsichtlich der Ausschöpfung der Notenskala als auch hinsichtlich der unterschiedlichen Bewertungen, die eine Einrichtung nach den verschiedenen Kriterien bekommen konnte, sehr differenziert. Befürchtungen, es könnte zu einer Noteninflation kommen, haben sich nicht bestätigt. Nach dem zentralen Kriterium „Forschungsqualität“ wurden nur 4 – 5 % aller Forschungseinheiten mit „exzellent“ bewertet.

Die Rückmeldungen aus den Einrichtungen und Nachfragen nach zusätzlichen Daten sprechen dafür, dass diese die Ergebnisse des Forschungsratings intensiv nutzen. Dabei ist insbesondere die häufige Frage nach den Ergebnissen vergleichbarer Forschungseinheiten an anderen Einrichtungen ein Indiz dafür, dass der Nutzen des Verfahrens durch eine Veröffentlichung der Ergebnisse aller Forschungseinheiten weiter gesteigert werden kann, die in der Pilotstudie aus Datenschutzgründen unterbleiben musste. Bei einer Wiederholung oder Ausweitung des Forschungsratings sollte deshalb die Personenbeziehbarkeit der Bewertungen durch entsprechende Regeln für den Zuschnitt der Forschungseinheiten ausgeschlossen werden. Wenn dies in bestimmten Fächern oder Einrichtungen nicht möglich ist, ist bereits zum Zeitpunkt der Erfassung der Forschungseinheiten die Zustimmung der Betroffenen zur Veröffentlichung der Bewertungsergebnisse einzuholen. Unter diesen Bedingungen können und sollten die Bewertungen der einzelnen Forschungseinheiten künftig veröffentlicht werden.

Neben den Bewertungsergebnissen haben die Einrichtungen auch die Datenberichte erhalten, die den Gutachtern als Bewertungsgrundlage vorgelegen haben. Diese enthalten mit den Perzentilwerten zu den quantitativen Indikatoren auch Angaben, aus denen die relative Position der jeweiligen Einrichtung in Deutschland abzulesen ist.[27] Die Nachfragen zu diesen Daten zeigen, dass das Forschungsrating in vielen

[27] vgl. Fn. 11, S. 16.

Einrichtungen das Bewusstsein für die Bedeutung einer kontinuierlichen Datensammlung gestärkt hat. Ein Grund dafür könnte sein, dass die Erhebung im Namen wissenschaftlicher Gutachter und nicht einer privaten Rankinginitiative erfolgte.

Neben dem unmittelbaren praktischen Nutzen der Ergebnisse für die Einrichtungen können auch die betroffenen Fächer aus dem Bewertungsverfahren dadurch produktive Anstöße erhalten, dass die im Forschungsrating verbindlich gemachten Qualitätskriterien zu einer nachhaltigen Diskussion ihrer fachlichen Ziele und Maßstäbe führen. Im Hinblick darauf bedürfen auch die normativen Effekte des Forschungsratings einer kritischen Dauerbeobachtung durch die Fachgemeinschaften. Es wäre wünschenswert, wenn die Ergebnisse dieser Diskussion die Weiterentwicklung des Forschungsratings befruchten würden. Im Hinblick auf die erwartbaren Lerneffekte scheint es nützlich, die für die Pilotstudien erhobenen Basisdaten für Sekundäranalysen, insbesondere für Vorhaben der Evaluationsforschung, verfügbar zu machen, wobei die Anonymisierung des Materials gewährleistet sein muss.

Einzelne Einrichtungen haben darauf aufmerksam gemacht, dass es für sie hilfreich wäre, sich z. B. in Hinblick auf Personaleinsatz, Drittmittel oder Publikationszahlen mit bestimmten anderen Einrichtungen im Sinne eines Benchmarks direkt vergleichen zu können. Dagegen spricht allerdings, dass die Daten für ein quantitatives Ranking verwendet werden könnten und dass ihre Veröffentlichung einen Legitimationsdruck auf die Gutachter erzeugen würde. Dies könnte zu zurückhaltenden, konformistischen Bewertungen führen und die Rekrutierung von Gutachtern erschweren. Der mögliche Nutzen einer Veröffentlichung ausgewählter Daten sollte unter Berücksichtigung dieser Bedenken im Dialog mit den Adressaten geprüft werden.

Das Forschungsrating nimmt eine retrospektive Bewertung bereits erbrachter Leistungen vor, soll aber auch eine Aussage über den Status Quo und das Entwicklungspotential der bewerteten Einheiten erlauben. Angesichts der Tatsache, dass die Datengrundlage in jedem Fall mehrere Jahre in die Vergangenheit zurückreicht, ist es wichtig, den Einrichtungsleitungen diese Steuerungsinformationen möglichst bald nach dem Stichtag zur Verfügung zu stellen. Die Ablaufplanung künftiger Forschungsratings sollte unter diesem Gesichtspunkt optimiert werden (vgl. II.1, S. 31 ff.).

Der Nutzen der Ergebnisse steigt, wenn durch eine erneute Bewertung erkennbar wird, ob die Bemühungen einzelner Einrichtungen zur Verbesserung der Forschungsleistungen Erfolg gehabt haben. Solche Effekte können jedoch erst einige Jahre nach der ersten Bewertung eintreten. Kürzere Bewertungsintervalle bringen die Gefahr einer Übersteuerung mit sich. Auch der Aspekt des Aufwands für Einrichtungen und Gutachter spricht für einen mehrjährigen Turnus. Auf der anderen Seite besteht bei großen Abständen die Gefahr, dass zu lange nach veralteten Informationen gehandelt wird. Ein Takt von etwa fünf bis sechs Jahren wäre auch nach internationalen Erfahrungen empfehlenswert.

### II.6. Zum Aufwand des Verfahrens

Der Aufwand der Pilotstudie war in Anbetracht der Neuartigkeit des Verfahrens angemessen. Das Forschungsrating ist grundsätzlich darauf angelegt, es auf andere Fächer zu übertragen und in den einzelnen Fächern turnusmäßig zu wiederholen. Dabei kann der Aufwand schrittweise reduziert werden. Die Spielräume für eine Reduzierung des Aufwands sind für die möglichen nächsten Schritte unterschiedlich:

<u>Übertragung auf andere Fächer</u>

Bei einer Bewertung weiterer, bisher nicht bewerteter Fächer wäre damit zu rechnen, dass der Aufwand je Fach sich zunächst in ähnlichen Dimensionen bewegt wie in der abgeschlossenen Pilotstudie. Der Zeitaufwand für die Gutachter wäre ähnlich, da für jedes neue Fach zunächst Indikatoren entwickelt werden müssten. Dabei könnten benachbarte Fächer mit ähnlichen Wissenschaftskulturen voneinander profitieren. Ein flexibles, mit geringerem Aufwand an weitere Fächer anpassbares System würde deshalb voraussichtlich im Laufe der Zeit entstehen, insbesondere dann, wenn auch Vorbilder aus den Geistes- und den Technikwissenschaften vorliegen.

Bisher nicht bewertete Fächer werden in den einzelnen Einrichtungen nicht auf die Datenerhebung vorbereitet sein, so dass sich ihr Aufwand, trotz der durch eine längere Vorlaufzeit erzielbaren Vereinfachung und möglicher Lernprozesse in den zentralen Verwaltungen, zunächst nicht wesentlich von dem unterscheiden wird, der für Chemie und Soziologie in der Pilotstudie erforderlich war.

Wiederholung in bereits bewerteten Fächern

Bei einer Wiederholung des Verfahrens in bereits einmal bewerteten Fächern würde demgegenüber zum einen die Indikatorenentwicklung deutlich kürzer ausfallen können. Darüber hinaus ist zu erwarten, dass bei der Datenerhebung eine Routinisierung eintritt. Entscheidend dafür ist, dass die wissenschaftlichen Einrichtungen Grunddaten zur Forschung in zentralen Datenbanken so sammeln, dass sie möglichst flexibel ausgegeben werden können. Da die für das Forschungsrating benötigten Daten nicht ungewöhnlich sind, sondern den internationalen Gepflogenheiten für eine Forschungsbewertung entsprechen, würde die Datenerhebung durch die an vielen wissenschaftlichen Einrichtungen geplante oder schon in Umsetzung befindliche Professionalisierung des Forschungscontrollings deutlich vereinfacht werden. Dieser Prozess sollte durch eine Abstimmung mit anderen Daten erhebenden Institutionen über die Erhebungsformate weiter unterstützt werden. Wenn zudem durch eine längere Vorlaufzeit eine retrospektive Erhebung entfiele oder zumindest in geringerem Maß notwendig wäre, würde der Aufwand seitens der bewerteten Einrichtungen erheblich sinken.

Beide Bewertungsgruppen haben betont, wie entscheidend für die Qualität der Begutachtung eine sorgfältige Datenbereinigung ist. Eine ausreichende personelle Kapazität für die administrative Unterstützung des Forschungsratings ist deshalb wichtig, um den Arbeitsaufwand für die Gutachter nicht in prohibitive Höhen zu treiben. In Fächern wie der Soziologie ist durch eine Verbesserung der Publikationsdaten, die die Gutachter von der Aufgabe entbindet, in zahlreichen Einzelfällen Kontrollen vorzunehmen bzw. in Auftrag zu geben, eine deutliche Reduzierung des Aufwands möglich. Die Zahl der zu bewertenden Forschungseinheiten könnte in Fächern wie der Soziologie dadurch verringert werden, dass die Regeln für die Definition dieser Einheiten zugunsten größerer Organisationseinheiten verändert werden. Dabei wäre allerdings zu bedenken, dass die Validität der Bewertungen aus fachlicher Sicht leiden würde, wenn die bewerteten Aggregate keine real für die Forschungsaktivitäten relevanten, handlungsfähigen Einheiten wären. Die Kosten für die Publikations- und Zitationsanalysen sind höher ausgefallen als erwartet, was bei Planungen künftiger Ratings zu berücksichtigen ist.

Der Ablauf der Bewertungsphase mit arbeitsteiliger Vorbewertung und anschließender plenarer Abstimmung ist sowohl bei einer Aufnahme neuer Fächer als auch bei

einer Wiederholung des Verfahrens beizubehalten. In Fächern, in denen es keine verlässlichen Indikatoren für die Rezeption wissenschaftlicher Arbeiten gibt, ist die direkte Begutachtung ausgewählter Publikationen durch Gutachter als valides Verfahren zur Qualitätsbeurteilung unersetzbar. Dies erhöht zwar den Bewertungsaufwand, ist aber aus fachlicher Sicht notwendig und angemessen.

Eine deutlich reibungslosere und effizientere Abwicklung der Datenerhebung ist durch eine Investition in ein modulares, an verschiedene Fächer anzupassendes Online-Erhebungsinstrument möglich. Die Entscheidung darüber, ein solches System zu entwickeln, sollte spätestens dann fallen, wenn das Forschungsrating in einem der in der vorliegenden Studie bereits bewerteten Fächer wiederholt wird.

Eine Reduzierung des Aufwands ist nicht mit einer Beschleunigung des Verfahrens zu verwechseln, die nicht möglich ist, ohne die Differenzierung der Bewertung und die für die Akzeptanz des Verfahrens entscheidende Berücksichtigung unterschiedlicher Fachkulturen aufzugeben; Forderungen der Einrichtungen nach längeren Vorlaufzeiten sprechen eher dafür, dass ein weniger dichtes Verfahren den Aufwand insgesamt verringern helfen würde.

Beim Vergleich des Aufwands für das Forschungsrating mit anderen Evaluations- und Bewertungsverfahren ist zu berücksichtigen, dass kein anderes Verfahren einen so differenzierten, nahezu flächendeckenden Vergleich der Institutionen des öffentlichen Forschungssektors leistet, der die Leistungen von 9.700 Wissenschaftlern (6.800 VZÄ) im Fall der Chemie und 1.400 Wissenschaftlern (1.160 VZÄ) im Fall der Soziologie erfasst. In Anbetracht von Umfang, Qualität und Differenziertheit der Bewertungen, die ein Forschungsrating leistet, sowie der Ziele, denen es dient, ist ein hoher Aufwand grundsätzlich gerechtfertigt. Aus Sicht der Steuerungsgruppe ist eine Weiterentwicklung des Verfahrens mit dem Ziel, den Aufwand zu reduzieren und zugleich die Validität und Verlässlichkeit der Bewertungen zu optimieren, sinnvoll und geboten.

**Anhang: Vergleichbarkeit mit veröffentlichten Rankings**

Unter den Alleinstellungsmerkmalen des in der Pilotstudie durchgeführten Forschungsratings (vgl. B.I, S. 27) sind einige, die zur Folge haben, dass seine Ergebnisse mit denen veröffentlichter Rankings nicht direkt vergleichbar sind. Zentrale Gründe dafür sind:

- Die Differenzierung in mehrere Bewertungskriterien;
- der Ausweis einer internen Differenzierung innerhalb einzelner Einrichtungen;
- die Berücksichtigung außeruniversitärer Einrichtungen.

Unterschiede zwischen den Ergebnissen verschiedener Bewertungsverfahren sind zudem schon deshalb zu erwarten, weil sich diese nur selten auf den selben Erhebungszeitraum beziehen. Darüber hinaus gibt es spezifische methodische Gründe, die bei Vergleichen der Ergebnisse des Forschungsratings mit bekannten Rankings berücksichtigt werden müssen.

Das Förder-Ranking der DFG beansprucht nicht, die Forschungsleistungen der analysierten Einrichtungen umfassend zu bewerten. Daten, wie sie im Förder-Ranking analysiert werden, gehen im Rahmen des vom Wissenschaftsrat entwickelten Forschungsratings zum Teil in die Bewertungsgrundlage ein. Bei einem Vergleich der von der DFG publizierten Drittmitteldaten mit denen, die im Forschungsrating verwendet werden, ist zu berücksichtigen:

- Das Förder-Ranking berichtet über bewilligte Mittel und fasst sie danach zusammen, welchem Fachkollegium und Fach das jeweilige Projekt zur Begutachtung zugeordnet wurde. Zahlen aus dem DFG-Förder-Ranking für die Chemie können deshalb im Einzelfall bspw. auch die von Chemikern begutachteten Projekte aus einem materialwissenschaftlichen Institut umfassen, hingegen die von Physikern begutachteten Projekte aus einem chemischen Institut ausschließen. Demgegenüber verwendet das Forschungsrating Daten über die von bestimmten, durch die jeweilige Einrichtung einem Fach zugeordneten Organisationseinheiten verausgabten Drittmittel.
- Die Kategorie „SOZ“ im DFG-Förder-Ranking umfasst die Sozial- und Verhaltenswissenschaften einschließlich Politikwissenschaften, Erziehungswissenschaften,

Wirtschaftswissenschaften und Psychologie. Eine Aussage über die Soziologie an einer Einrichtung ist auf dieser Basis nicht möglich.

Bei einem Vergleich mit den jeweils aktuellen Forschungsrankings des Centrums für Hochschulentwicklung (CHE) für die bewerteten Fächer sind unter anderem folgende methodische Differenzen zu berücksichtigen:

- Das CHE-Forschungsranking bezieht sich auf Organisationseinheiten, die an den Studiengängen eines Fachs beteiligt sind. Demgegenüber wurden für die Pilotstudie alle im Fach forschungsaktiven Einheiten erfasst. Dies hat besonders für die Soziologie eine deutliche Ausweitung des Gegenstandsbereichs gegenüber dem CHE-Forschungsranking zur Folge.
- Diskrepanzen in den relativen Zahlen, insbesondere in den Promotionsquoten, sind zum Teil mit tatsächlichen Veränderungen der Personalzahlen zwischen den Erhebungszeiträumen von CHE und Pilotstudie zu erklären.
- Die Publikationszahlen für die Soziologie lassen sich nicht unmittelbar vergleichen, da das CHE nur gewichtete Publikationszahlen veröffentlicht. Bedeutsam ist jedoch, dass die Ergebnisse der Publikationsrecherche in der Pilotstudie Forschungsrating durch die betroffenen Wissenschaftler kontrolliert wurden. Aufgrund ihrer Korrekturen wurde die Anzahl der berücksichtigten Dokumente von ca. 5.300 auf 10.600 verdoppelt. Ohne umfassende Mitwirkung der betroffenen Wissenschaftler ist in der Soziologie eine erhebliche Untererfassung der Publikationen nach den Erfahrungen aus der Pilotstudie unvermeidlich und erklärt die deutlich abweichenden Ergebnisse der Publikationsanalyse des CHE.
- Der Erhebungszeitraum des CHE ist kürzer als der des Forschungsratings. Damit ist auch das maximale Zitationszeitfenster für Zitationsanalysen geringer. Im übrigen ist darauf hinzuweisen, dass das CHE bislang keine subfieldnormierten Zitationsdaten verwendet.
- Zur Identifikation sogenannter „forschungsstarker“ Universitäten verrechnet das CHE absolute und relative Publikations-, Drittmittel- und Promotionsdaten. Diese Indikatoren sind im Forschungsrating unterschiedlichen Kriterien zugeordnet. Bspw. weist das CHE keine separate Effizienzbewertung aus, sondern bezieht die relativen Daten in die Gesamtbewertung ein.

Die übrigen, in Zeitschriften veröffentlichten Rankings sind noch schwieriger mit dem Forschungsrating zu vergleichen, da die Methodik dieser Rankings in der Regel intransparent ist. Beispielsweise ist im Ranking der Zeitschrift „Fokus“ zwar ein fachspezifischer „Forschungs“-Wert für jede Universität ausgewiesen. Dieser beruht auf Daten für die Drittmitteleinwerbung, den Promotionsquoten sowie einer Reputationsbefragung bei Professoren. Eine Erklärung dafür, warum bspw. einige im Forschungsrating Soziologie mit sehr gutem Impact abschneidenden Einrichtungen beim Fokus-Ranking 2007 in der Schlussgruppe platziert sind, ist jedoch ohne Kenntnis der Daten nicht möglich.

**Anlagen**

Pilotstudie Forschungsrating Chemie: Abschlussbericht der Bewertungsgruppe (Drs. 8370-08)

Pilotstudie Forschungsrating Soziologie: Abschlussbericht der Bewertungsgruppe (Drs. 8422-08)