WISSENSCHAFTSRAT

GESCHÄFTSSTELLE

Drs. 8370-08
Köln, 11.01.2008

Pilotstudie Forschungsrating Chemie
Abschlussbericht der Bewertungsgruppe

Inhalt Seite

Vorbemerkung

Der Wissenschaftsrat hat im Juli 2005 beschlossen, das von ihm in seinen „Empfehlungen zu Rankings im Wissenschaftssystem“ vom November 2004[1] entworfene Verfahren eines Forschungsratings in einer Pilotstudie in den Fächern Chemie und Soziologie zu erproben. Die vom Wissenschaftsrat mit der Durchführung der Pilotstudie beauftragte Steuerungsgruppe hat auf Basis von Vorschlägen der großen Wissenschaftsorganisationen, der Gesellschaft Deutscher Chemiker und des Verbands der Chemischen Industrie die Mitglieder einer Bewertungsgruppe für die Chemie berufen. Diese Bewertungsgruppe hat das Verfahren an die Bedürfnisse der Chemie angepasst und auf Basis der erhobenen Daten die Forschungsleistungen deutscher Universitäten und außeruniversitärer Forschungseinrichtungen in der Chemie vergleichend bewertet. Die Ergebnisse dieser Bewertungen sind von der Steuerungsgruppe im Dezember 2007 veröffentlicht worden.[2]

Im vorliegenden Abschlussbericht fasst die Bewertungsgruppe Chemie ihre Erfahrungen aus der Pilotstudie zusammen und gibt Empfehlungen dazu ab, wie das Verfahren optimiert werden sollte, falls der Beschluss gefällt wird, es nach einigen Jahren erneut, ggf. auch in den Nachbardisziplinen der Chemie, durchzuführen.

Der vorliegende Abschlussbericht wurde von der Bewertungsgruppe Chemie am 11. Januar 2008 verabschiedet.

---

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

2 Steuerungsgruppe der Pilotstudie Forschungsrating im Auftrag des Wissenschaftsrates: Forschungsleistungen deutscher Universitäten und außeruniversitärer Forschungseinrichtungen in der Chemie. Ergebnisse der Pilotstudie Forschungsrating des Wissenschaftsrates. Köln 18.12.2007.

Kurzfassung

Der Wissenschaftsrat hat im Juli 2005 beschlossen, das im November 2004 empfohlene Verfahren zu einem Forschungsrating[3] zunächst in einer Pilotstudie in zwei Fächern, Chemie und Soziologie, zu erproben. Ziel der Pilotstudie war es auszuloten, inwiefern der in den Empfehlungen vorgegebene Rahmen eine Operationalisierung für verschiedene Fächer zulässt und dann praktisch umsetzbar ist. Die für die fachliche Umsetzung in der Chemie eingesetzte Bewertungsgruppe hat zunächst das Verfahren für ihr Fach angepasst und dann die vergleichende Bewertung der Forschungsleistung deutscher Universitäten und außeruniversitärer Einrichtungen vorgenommen, deren Ergebnisse am 18. Dezember 2007 veröffentlicht wurden.[4]

Die Bewertungsgruppe Chemie ist überzeugt, dass das gewählte Verfahren einer Gutachterbewertung auf der Basis zahlreicher quantitativer und qualitativer Indikatoren einfacheren Verfahren, etwa einer reinen Indikatorenauswertung, vorzuziehen ist. Der höhere Aufwand ist durch den großen Nutzen einer differenzierteren und belastbaren Bewertung gerechtfertigt. Die Bewertungsgruppe Chemie rät nachdrücklich davon ab, quantitative Indikatoren ohne eine qualitative Einordnung zum alleinigen Maß für längerfristige Entwicklungsplanungen zu machen. Erst eine Zusammenstellung quantitativer und qualitativer Aspekte und deren Abwägung durch Fachgutachter ermöglicht eine fundierte Bewertung. Sie sollte nicht unmittelbar an staatliche Mittelzuweisungen gekoppelt werden, sondern den betroffenen Einrichtungen in Verbindung mit anderen Instrumenten der Qualitätssicherung als Grundlage für ihre strategischen Planungen dienen.

Zu den Vorzügen des Verfahrens zählt seine Mehrdimensionalität. Diese ermöglicht es, Einrichtungen mit unterschiedlichen Missionen nach gleichen Maßstäben zu bewerten und ihnen ein individuelles Profil ihrer Stärken und Schwächen aufzuzeigen. Die Bewertungsgruppe Chemie hält die den drei Dimensionen „Forschung“, „Nachwuchsförderung“ und „Wissenstransfer“ zugeordneten Kriterien Forschungsqualität, Impact/Effektivität, Effizienz, Nachwuchsförderung, Transfer in andere gesellschaftliche Bereiche und Wissensvermittlung und -verbreitung grundsätzlich für geeignete Kriterien zur vergleichenden Bewertung der Leistungsfähigkeit in der Forschung.

---

3 Wissenschaftsrat: „Empfehlungen zu Rankings im Wissenschaftssystem. Teil 1: Forschung.“ in Empfehlungen und Stellungnahmen 2004, Bd. I. Köln 2005. 159 – 220.

4 Vgl. Steuerungsgruppe der Pilotstudie Forschungsrating im Auftrag des Wissenschaftsrates: Forschungsleistungen deutscher Universitäten und außeruniversitärer Forschungseinrichtungen in der Chemie. Ergebnisse der Pilotstudie Forschungsrating des Wissenschaftsrates. Köln 18.12.2007.

Um die Verlässlichkeit der Bewertung weiter zu erhöhen und zugleich den Aufwand zu begrenzen, schlägt die Bewertungsgruppe eine Reihe von Verfahrensänderungen für den Fall vor, dass die Chemie in einigen Jahren erneut bewertet werden soll:

- In Einzelfällen war die Heterogenität der „Forschungseinheiten“, auf deren Ebene – unterhalb der Einrichtungsebene angesiedelt – die Forschungsqualität bewertet wurde, problematisch. Hier ist eine stärkere Abstimmung bei der Bildung der Forschungseinheiten zu empfehlen.
- Die Bewertung von Forschung an den Disziplinengrenzen ist problematisch, wenn nur eine Disziplin isoliert bewertet wird. Hier ist zu empfehlen, dass künftig vor allem benachbarte Disziplinen parallel bewertet werden. Dann könnten solche Forschungseinheiten, die nicht eindeutig nur einer Disziplin zuzuordnen sind, ggf. von zwei Gutachtergruppen unterschiedlicher Fächer bewertet werden.
- Für die Bewertung der Effizienz sind konsistente Daten über die eingesetzten Finanzmittel wünschenswert. Hier mehr Transparenz und Vergleichbarkeit zu schaffen, ist ein dringendes politisches wie volkswirtschaftliches Desiderat.
- Die Kriterien „Transfer in andere gesellschaftliche Bereiche“ und „Wissensvermittlung und -verbreitung“ waren nicht deutlich genug voneinander abgegrenzt, was sowohl für die Erhebung der zugehörigen Indikatoren als auch für den Bewertungsprozess negative Konsequenzen hatte. Die Bewertungsgruppe empfiehlt für die Chemie eine stärkere Differenzierung in „Transfer in die Wirtschaft“ und „Wissensvermittlung in die Öffentlichkeit“ und schlägt für letztgenanntes Kriterium eine modifizierte Datengrundlage vor.
- Die in der Pilotstudie erhobenen Indikatoren sind nach Ansicht der Bewertungsgruppe im Wesentlichen beizubehalten. An einigen Stellen sind aber, vor allem zur Verringerung des Erhebungsaufwandes, Modifikationen notwendig. Auch eine anwenderfreundliche Vereinfachung und Professionalisierung des Erhebungsinstruments wird dringend empfohlen.
- Der in Einzelfällen sehr hohe Erhebungsaufwand lässt darauf schließen, dass die Datenvorhaltung der teilnehmenden Einrichtungen den Selbststeuerungsanforderungen an zunehmend autonome Hochschulen noch nicht überall gerecht wird. Eine Verstetigung, verbunden mit längeren Vorlaufzeiten, lässt Effizienzgewinne bei der Datenerhebung aufseiten der Einrichtungen erwarten.

Die im Abschlussbericht der Bewertungsgruppe näher ausgeführten Vorschläge zur Optimierung des Verfahrens sind wichtige Lehren, die bei einer möglichen Wiederholung und Ausweitung des Verfahrens berücksichtigt werden sollten. Die Bewertungs-

gruppe Chemie hält das für das Fach Chemie erprobte Verfahren grundsätzlich für gut geeignet, die Forschungsleistung der deutschen universitären und außeruniversitären Forschung im Fach Chemie vergleichend zu bewerten. Sie begrüßt ausdrücklich die durch ein solches Verfahren geschaffene Transparenz und das differenzierte Aufzeigen spezifischer Stärken und Schwächen einzelner Einrichtungen nach verschiedenen Kriterien, das vor allem den Einrichtungen selbst wichtige Erkenntnisse liefern kann und ihnen eine Basis bietet, ihre Profile weiterzuentwickeln. Bei der Interpretation der Ergebnisse der vorliegenden Untersuchung ist zu beachten, dass es sich um eine Pilotstudie handelte: Der Erkenntniswert des Forschungsratings würde sich durch eine Wiederholung des Verfahrens voll entfalten, da erst dann Trends sichtbar gemacht und Veränderungen erfasst werden können. Daher spricht sich die Bewertungsgruppe dafür aus, das Verfahren unter Berücksichtigung der genannten Modifikationen in einem gewissen zeitlichen Abstand zu wiederholen und möglichst auch auf die Nachbardisziplinen der Chemie, Biologie und Physik, auszuweiten.

A. Ausgangslage

Die Durchführung der Pilotstudie Forschungsrating geht zurück auf die „Empfehlungen zu Rankings im Wissenschaftssystem. Teil 1: Forschung“, die der Wissenschaftsrat im November 2004 verabschiedet hat. Darin empfiehlt der Wissenschaftsrat, die Forschungsleistungen von Universitäten und außeruniversitären Einrichtungen vergleichend zu bewerten, um die Einrichtungen bei strategischen Entscheidungen zu unterstützen und durch mehr Transparenz den Wettbewerb zu fördern. Der Wissenschaftsrat hat sich in diesen Empfehlungen kritisch mit bestehenden Ranking-Verfahren auseinandergesetzt und angesichts der internationalen Erfahrungen mit vergleichenden Bewertungsverfahren eine ausschließlich auf quantitativen Indikatoren basierende Forschungsbewertung ebenso ausgeschlossen wie eine reine Reputationsmessung. Das von ihm vorgeschlagene Forschungsrating ist demgegenüber gekennzeichnet durch:

- das Prinzip des „Informed Peer Review“, d.h. eine Bewertung durch Gutachter[5] auf der Basis standardisierter, quantitativer und qualitativer Daten, die fachspezifisch bestimmt werden;
- Mehrdimensionalität, d. h. Unterscheidung verschiedener Leistungskriterien, die nicht zu einer Gesamtbewertung verrechnet werden, um so unterschiedlichen Aufgaben verschiedener Arten von Einrichtungen gerecht zu werden;
- Verzicht auf Ranglistenbildung, um irreführende Scheingenauigkeit und dadurch erzeugte Fehlsteuerungseffekte zu vermeiden.

In den Empfehlungen war vorgesehen, das vorgeschlagene Verfahren möglichst in einer Pilotstudie zu erproben. Im Juli 2005 wurde beschlossen, eine solche Pilotstudie in den Fächern Chemie und Soziologie durchzuführen.[6] Der Beschluss, in der Pilotstudie das Fach Chemie zu berücksichtigen, wurde unter anderem durch einen Vorschlag der Gesellschaft Deutscher Chemiker (GDCh) und des Verbandes der Chemischen Industrie (VCI) angeregt, die die Pilotstudie Chemie in der Folgezeit unterstützt haben.

---

[5] 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.

[6] Die folgenden Ausführungen beziehen sich – sofern nicht anders ausgewiesen – immer nur auf die Pilotstudie im Fach Chemie.

A.I. Organisation und Ablauf der Pilotstudie

Für die Durchführung der Pilotstudie ist eine vom Wissenschaftsrat eingesetzte Steuerungsgruppe verantwortlich. In ihr sind neben Mitgliedern der Wissenschaftlichen Kommission des Wissenschaftsrates und weiteren Sachverständigen die großen Wissenschaftsorganisationen Deutsche Forschungsgemeinschaft (DFG), Fraunhofer-Gesellschaft (FhG), Helmholtz-Gemeinschaft Deutscher Forschungszentren (HGF), Hochschulrektorenkonferenz (HRK), Max-Planck-Gesellschaft (MPG) und Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz (WGL) durch ex officio-Mitglieder, in der Regel durch Vizepräsidenten, vertreten. Sechs Länder, der Bund und der VCI sowie die Geschäftsstellen der großen Wissenschaftsorganisationen entsenden Gäste in die Steuerungsgruppe.

Für die Operationalisierung des Verfahrens und die Bewertung der Forschungsleistung im Fach Chemie wurde eine Bewertungsgruppe eingesetzt, die sich aus insgesamt 15 Fachvertretern zusammensetzte. Die Mitglieder der Bewertungsgruppe Chemie sind von der Steuerungsgruppe aus Vorschlägen der großen Wissenschaftsorganisationen sowie der GDCh und des VCI ausgewählt worden. Bei der Auswahl der Gutachter wurde darauf geachtet, möglichst viele Teilgebiete der Chemie durch die Expertise der Bewertungsgruppe abzudecken. Zugleich sollten auch ausländische Gutachter beteiligt werden, die allerdings in der Lage sein mussten, in deutscher Sprache verfasste Unterlagen zu lesen. Die gewünschte Internationalität konnte auch über langjährige internationale Erfahrung der Gutachter erreicht werden. Zudem war die Steuerungsgruppe durch einen Gast in der Bewertungsgruppe vertreten, um die Verfahrenskohärenz sicherzustellen.

Die Pilotstudie Forschungsrating ist in vier Phasen unterteilt: 1. Fachspezifische Operationalisierung durch die Bewertungsgruppe, 2. Erhebung der Daten bei den Einrichtungen über Fachkoordinatoren, 3. Bewertung der Daten durch die Bewertungsgruppe, 4. Berichte und Empfehlungen zum Verfahren.

A.II. Operationalisierung des Forschungsratings für das Fach Chemie

Das vom Wissenschaftsrat empfohlene Forschungsrating sieht eine Anpassung des Verfahrens an das zu bewertende Fach vor. Die drei Leistungsdimensionen Forschung, Nachwuchsförderung und Wissenstransfer bilden dafür den fächerübergreifend konstant zu haltenden Rahmen.

II.1. Definition des Fachs Chemie

Die Erfassung der im Bereich der Chemie forschungsaktiven Einrichtungen setzte zunächst eine Definition des Fachs voraus. Die Bewertungsgruppe bestimmte in Abstimmung mit der Steuerungsgruppe zehn Teilgebiete, die zusammen das Gebiet der Chemie definieren und gegen andere Fächer abgrenzen. Dabei ging sie zunächst von der ab 2003 geltenden Benennung der DFG-Fachkollegien aus, entschied sich jedoch letztlich, davon abzuweichen, da die Institute an den Universitäten in der Regel noch den traditionelleren Teilgebieten entsprechen und somit eine Zuordnung auf Basis dieser klassischen Gebietsbezeichnungen leichter fallen sollte. Aufgrund dieser Überlegungen hat die Bewertungsgruppe folgende Teilgebiete definiert: Analytische Chemie, Anorganische Chemie, Biochemie und biologische Chemie, Organische Chemie, Lebensmittelchemie, Medizinische/Pharmazeutische Chemie, Physikalische Chemie, Polymerchemie, Technische Chemie und Theoretische Chemie. Bei einigen Teilgebieten war eine eindeutige Zuordnung zur Chemie schwierig, vor allem in der Biochemie und der Medizinischen/Pharmazeutischen Chemie.[7] In Grenzfällen wurden die Einrichtungen aufgefordert, selbst zu entscheiden, ob die betreffenden Forschungseinheiten vorrangig der Chemie zugeordnet werden können.

II.2. Definition der Kriterien und Skalen

Die Definition der Bewertungskriterien beruht auf den Empfehlungen des Wissenschaftsrates von November 2004. Darin wurden drei Dimensionen der Forschungsleistung (Forschung, Nachwuchsförderung und Wissenstransfer) vorgegeben und deren Bewertung nach neun Kriterien vorgeschlagen. Die genaue Ausgestaltung der Bewertungskriterien durch Bewertungsaspekte sowie die Zuordnung von Indikatoren soll jeweils fachspezifisch erfolgen. Grundsätzlich müssen die Kriterien dazu geeignet sein, alle Einrichtungen in einem Fach nach einheitlichen Maßstäben bewerten zu können. Darüber hinaus ist eine fächerübergreifende Standardisierung der Kriterien wünschenswert.

Die Bewertungsgruppe Chemie hat in Abstimmung mit der Steuerungsgruppe die Bewertungskriterien von neun auf sechs Kriterien reduziert. Die Reduktion betraf die Dimensionen Wissenstransfer und Nachwuchsförderung, die nunmehr aus zwei bzw. einem Bewertungskriterium bestehen, während die Dimension Forschung weiterhin durch drei Kriterien definiert ist. Hauptgrund für diese Vereinfachung war, dass in den

[7] Das Fachgebiet der (chemischen) Verfahrenstechnik ist gänzlich ausgeklammert, da es in Deutschland typischerweise nicht der Chemie zugeordnet ist.

betroffenen Dimensionen nicht genügend etablierte und zuverlässig zu erhebende Indikatoren verfügbar sind, um eine nach mehreren Kriterien differenzierende Bewertung zu ermöglichen. Zudem war aus Sicht der Gutachter zu befürchten, dass eine hohe Zahl von Kriterien den Eindruck erwecken könnte, die Transferdimension sei wichtiger als die Dimension Forschung.

Tabelle 1: Dimensionen und Kriterien der Bewertung

Dimension	Kriterium
Forschung	I. Forschungsqualität (Ebene Forschungseinheit)
	II. Impact/Effektivität[1)]
	III. Effizienz
Nachwuchsförderung	IV. Nachwuchsförderung
Wissenstransfer	V. Transfer in andere gesellschaftliche Bereiche
	VI. Wissensvermittlung und -verbreitung

1) Unter „Impact/Effektivität“ ist die Sichtbarkeit der Forschung in der scientific community zu verstehen.

Die Bewertungsgruppen haben in Abstimmung mit der Steuerungsgruppe auch die Bewertungsskala definiert. Abweichend von der ursprünglichen Empfehlung des Wissenschaftsrates hatte die Steuerungsgruppe zu Beginn der Pilotstudie eine fünfstufige Skala mit den Werten 5=„exzellent“, 4=„sehr gut“, 3=„gut“, 2=„befriedigend“ und 1=„nicht befriedigend“ vorgeschlagen. Die Bewertungsgruppe hat diese Skala übernommen und, in Absprache mit der Bewertungsgruppe Soziologie, zudem die oberste Stufe „exzellent“ zusätzlich verbal definiert als „gehört nach internationalen Maßstäben zu den führenden Forschungseinheiten (Forschungsqualität) bzw. Einrichtungen (Impact / Effektivität)“ oder „ist eine führende Einrichtung in Deutschland“ (übrige Kriterien).

Im Bewertungskriterium Wissensvermittlung und -verbreitung wurde die Skala auf drei Stufen („unterdurchschnittlich“, „durchschnittlich“, „überdurchschnittlich“) reduziert, weil die Datengrundlage zu diesem Kriterium zu heterogen und zu schmal war und nur einen quantitativen Indikator umfasste. Demgegenüber wurde für das auf Ebene von Forschungseinheiten bewertete Kriterium Forschungsqualität zusätzlich die Stufe „sehr gut bis exzellent“ eingeführt, da die Datengrundlage hier sehr gut war und – aufgrund der Asymmetrie der Verteilungen mehrerer Indikatoren – gerade im oberen Notenbereich eine feinere Differenzierung ermöglichte.

II.3. Zuordnung von Indikatoren

Die sechs Bewertungskriterien wurden in einzelne fachspezifische Bewertungsaspekte unterteilt. Diesen wiederum wurden jeweils bestimmte qualitative und quantitative Daten als Bewertungsgrundlage zugeordnet. Der Zusammenhang von Dimensionen, Kriterien, Aspekten und Daten ist in einer sogenannten Bewertungsmatrix wiedergegeben.[8]

Die Auswahl der Indikatoren nahm die Bewertungsgruppe in Abstimmung mit der Steuerungsgruppe vor. Für die Auswahl der Indikatoren waren mehrere Faktoren entscheidend: die Erhebbarkeit und Belastbarkeit der Daten und die unabhängige Aussagekraft eines Indikators zwecks Verminderung von Redundanz. Zu jedem Kriterium wurden möglichst sowohl quantitative als auch qualitative Indikatoren ausgewählt.

Eine erste Fassung der Bewertungsmatrix wurde in einem Pretest erprobt. An diesem Pretest nahmen eine Universität und drei außeruniversitäre Einrichtungen teil. Dadurch konnte überprüft werden, ob der Satz der Indikatoren den verschiedenen Arten von Einrichtungen mit ihren verschiedenen Aufgaben gerecht wird. Die teilnehmenden Einrichtungen wurden um Feedback gebeten, der in die Entwicklung der endgültigen Bewertungsmatrix und Fragebögen einging. Der Pretest lieferte eine wichtige Grundlage für die abschließende Anpassung der Indikatoren.

II.4. Definition der Forschungseinheiten

Im Zuge der Vorbereitung der Pilotstudie hat die Steuerungsgruppe beschlossen, bei der Bewertung der Forschungsqualität auch Unterschiede innerhalb einer Einrichtung wiederzugeben. Dafür wurde die unterhalb der Ebene Einrichtung angesiedelte Erhebungs- und Bewertungsebene der „Forschungseinheit“ eingeführt.

Eine Forschungseinheit ist im Rahmen der Pilotstudie Chemie als eine Gruppe von mindestens drei hauptamtlichen Wissenschaftlern definiert, die über einen längeren Zeitraum ein zusammenhängendes Forschungsprogramm verfolgt und in der Regel mit vorhandenen Abteilungen, Instituten, Zentren oder anderen Organisationseinheiten identisch ist. Es wurde empfohlen, den Zuschnitt der Forschungseinheiten an den Teilgebieten der Chemie zu orientieren und für eine Einrichtung mittlerer Größe ca. drei bis sechs Forschungseinheiten zu definieren.

---

[8] Die Bewertungsmatrix ist im Internet veröffentlicht unter http://www.wissenschaftsrat.de/texte/Bewertungsmatrix_Chem.pdf.

Um der Tatsache Rechnung zu tragen, dass Forschergruppen häufig über mehrere Einrichtungen verbunden kooperativ forschen, konnten auch „institutionenübergreifende Forschungseinheiten“ benannt werden. In diesen Fällen wurden die Publikationsleistungen von gemeinsam berufenen Wissenschaftlern beiden Forschungseinheiten voll zugerechnet. Die übrigen Forschungsleistungen sollten die Einrichtungen nach Möglichkeit getrennt aufführen.

II.5. Entwicklung von Fragebögen

Da nur ein kleiner Teil der nach der Bewertungsmatrix (s. II.3, S. 11) notwendigen Daten in vorhandenen Datenbanken erfasst ist und diese Datenbanken zudem untereinander inkompatible Definitionen von Fachgebieten und/oder Organisationseinheiten verwenden, mussten fast alle Daten für die Pilotstudie neu erhoben werden. Bei der Entwicklung der dafür notwendigen Fragebögen wurde die Bewertungsgruppe durch eine Unterarbeitsgruppe der Steuerungsgruppe unterstützt, die aus vorwiegend administrativen Vertretern von Universitäten und außeruniversitären Instituten bestand. Zusätzlich wurde die Geschäftsstelle durch das Institut für Forschungsinformation und Qualitätssicherung (iFQ) und das Zentrum für Umfragen, Methoden und Analysen (ZUMA) beraten. Die inhaltliche Basis der Fragebögen bildete die zuvor erstellte Bewertungsmatrix. Insgesamt wurden für die Erhebung drei Fragebögen verwendet: Fragebogen I diente der Erfassung der Struktur der Einrichtungen (Forschungseinheiten, leitende Wissenschaftler), Fragebogen II erfasste die Daten auf Ebene der Einrichtung, Fragebogen III erfasste die Daten auf Ebene der Forschungseinheiten (zur Datenerhebung s. A.III).

Jeder Fragebogen bestand aus einem Text- und einem Tabellenteil, die mit handelsüblicher Office-Software zu bearbeiten waren. Im Textteil wurden Rahmeninformationen zum Profil/zur Mission der Einrichtungen bzw. Forschungseinheiten, Hintergrundinformationen und Selbstbeschreibungen abgefragt. Im Tabellenteil wurden Listen und quantitative Daten erhoben. Vorgaben hinsichtlich der Maximallänge oder -anzahl von Einträgen wurden durch Beschränkungen von Zeichen- bzw. Zeilenzahlen umgesetzt. Für eine Office-basierte Lösung sprach vor allem, dass eine aufwendige Zugangsdatenverwaltung innerhalb der einzelnen Einrichtungen unterbleiben konnte, da diese Fragebögen per e-mail verschickt werden konnten.

II.6. Rahmenbedingungen der Publikations- und Zitationsanalyse

Im Fach Chemie sind Publikationen in referierten Zeitschriften und Zitationen auf diese Publikationen als Indikatoren der Forschungsleistung schon lange etabliert und als Qualitätsstandards weitgehend anerkannt. Die verfügbaren Datenbanken erlauben eine belastbare Erfassung und Analyse der Publikationen und Zitationen für dieses Fach. Deshalb wurde beschlossen, eine ausführliche Publikations- und Zitationsanalyse für die Chemie in Deutschland durchzuführen und die dadurch generierten bibliometrischen Indikatoren in die Bewertung einzubeziehen.

Mit der Erstellung der Publikations- und Zitationsanalyse wurde nach beschränkter Ausschreibung das Institut für Wissenschafts- und Technikforschung (IWT) der Universität Bielefeld beauftragt. Datenbasis war das „Web of Science“ (WoS) des kommerziellen Anbieters Thomson Scientific (ehem. ISI).

In einem ersten Schritt wurden Publikationslisten für die einzelnen Einrichtungen erstellt, wobei die Recherche jeweils die Namen der von diesen benannten leitenden Wissenschaftler mit Adressbestandteilen der Einrichtung und dem Erhebungszeitraum (2001-2005) kombinierte. Die so erstellten Listen wurden dann an die Einrichtungen zurückgekoppelt und daraufhin ggf. korrigiert. Die korrigierten Publikationslisten waren Grundlage der umfangreichen Publikations- und Zitationsanalyse (vgl. III.4, S. 16 f.).

A.III. Datenerhebung

III.1. Organisation und Ablauf

Vor Einleitung der Datenerhebung musste festgelegt werden, auf welche Gegenstände sich die Bewertungen beziehen sollten. Dazu wurden in einem ersten Schritt alle staatlichen und ausgewählte private Universitäten sowie die von Bund und Ländern gemeinsam geförderten außeruniversitären Einrichtungen, die chemische Forschung durchführen, eingeladen, an der Pilotstudie Forschungsrating teilzunehmen. Insgesamt reagierten 78 Einrichtungen positiv auf diese Anfrage, eine Einrichtung zog ihre Teilnahme nachträglich zurück, so dass die Pilotstudie sich schließlich auf 77 Einrichtungen bezog. Mit der Einladung wurde auch darum gebeten, einen Fachkoordinator zu benennen, der in den Einrichtungen für die Koordinierung und Organisation der Datenerhebung zuständig sein sollte. Weil es neben der reinen Datener-

hebung auch um fachlich-strategische Fragen ging, wurde empfohlen, dafür einen Wissenschaftler zu benennen, der durch die Verwaltung unterstützt werden sollte.

Die Erhebung erfolgte in zwei Phasen: In der ersten Erhebungsphase wurden die Struktur und das leitende wissenschaftliche Personal der Chemie an den teilnehmenden Einrichtungen erfasst. In der anschließenden zweiten Phase wurden die bewertungsrelevanten Daten erhoben. So konnte nach Abschluss der ersten Phase parallel zur zweiten Phase bereits mit der Publikationsrecherche begonnen werden, die auf den Angaben zum leitenden wissenschaftlichen Personal basierte.

III.2. Erfassung der Forschungseinheiten

Die erste Erhebungsphase fand im Sommer 2006 statt. Die teilnehmenden Einrichtungen meldeten die leitenden Wissenschaftler der Chemie und ordneten sie Forschungseinheiten zu. Als „leitende Wissenschaftler“, die namentlich und unter Angabe des Zeitraums ihrer Beschäftigung benannt werden sollten, galten dabei Professoren bzw. Direktoren und Gruppenleiter mit Forschungsaufgaben.[9] Ausschlaggebend für deren Meldung war der Beschäftigungszeitraum, der sich mit dem Erhebungszeitraum überschneiden musste.

Zum Zuschnitt der Forschungseinheiten gab es Richtlinien, aber keine strikten Vorgaben, da eine einheitliche Struktur an allen chemisch forschenden Einrichtungen nicht vorausgesetzt werden konnte. Diese Offenheit bereitete in der ersten Erhebungsphase teilweise Probleme. In einigen Fällen wurden Forschungseinheiten benannt, die im Rückblick als zu groß bzw. zu heterogen oder aber zu klein und damit unterkritisch zu bezeichnen sind. Insgesamt wurde der Zuschnitt der Forschungseinheiten sehr individuell gehandhabt. Mögliche Gründe für diese Heterogenität sind:

- Die Einrichtungen verfügen in der Regel über gewachsene (Instituts-) Strukturen, die sich an der Lehre, nicht an der Forschung orientieren;
- chemische Forschung wird häufig in interdisziplinären Zusammenhängen und schwierig aufzuteilenden einrichtungsübergreifenden Kooperationen betrieben;
- in einigen Einrichtungen wird die Forschung von Service-Abteilungen unterstützt, deren eindeutige Zuordnung vor allem bei der Bewertung der Effizienz einer Einrichtung schwierig ist.

[9] Die Leitungsfunktion eines „leitenden Wissenschaftlers“ bezieht sich ausschließlich auf die Forschung, nicht auf administrative Funktionen. Emeriti, Vertretungs-, Gast- und Honorarprofessoren, Privatdozenten ohne Anstellungsvertrag sowie wissenschaftliche Mitarbeiter, wissenschaftliche und studentische Hilfskräfte sind nicht berücksichtigt.

Insgesamt meldeten die 77 teilnehmenden Einrichtungen 349 Forschungseinheiten. Das heißt, im Mittel hat jede Einrichtung etwa 4,5 Forschungseinheiten gebildet. Jeder Forschungseinheit wurden im Mittel etwa sechs leitende Wissenschaftler, darunter 3 Professoren/Direktoren zugeordnet. Mit 1038 Professoren zum Stichtag (davon nur 73 Professorinnen, d.h. 7%) haben deutlich mehr Wissenschaftler an der Pilotstudie Chemie teilgenommen als das Statistische Bundesamt dem Lehr- und Forschungsbereich „Chemie“ zuordnet (895 im Jahr 2005 – ohne FH-Professoren).[10] Insgesamt wurde also eine hohe Erfassungsquote erzielt.[11]

III.3. Datenerhebung

Die zweite Phase der Datenerhebung setzte im Oktober 2006 ein und war zunächst bis Mitte Dezember veranschlagt, lief nach Fristverlängerung aber bis Ende Januar 2007. In dieser Phase wurden mittels der zuvor entwickelten Fragebögen (vgl. II.5, S. 12 f.) bei den Einrichtungen und Forschungseinheiten die bewertungsrelevanten Daten erhoben. Die Fragebögen wurden in elektronischer Form versandt und konnten durch die Fachkoordinatoren auf Datenträgern oder per E-Mail an die Geschäftsstelle zurückgeschickt werden. Jede Einrichtung erhielt einen Fragebogen II, in dem die auf die Ebene der Einrichtung bezogenen Angaben abgefragt wurden, und je Forschungseinheit einen Fragebogen III für auf die Forschungseinheiten bezogene Angaben.

Neben den bei den Einrichtungen selbst erhobenen Daten wurden externe Daten aus mehreren Quellen verwendet: Die GDCh stellte Promotionszahlen der Universitäten bereit, der Fonds der Chemischen Industrie (FCI) meldete seine jährlich vergebenen Forschungspreise und die Alexander von Humboldt-Stiftung (AvH) stellte Angaben zu Gastwissenschaftlern zur Verfügung. Außerdem wurden die für die Publikations- und Zitationsanalyse notwendigen Daten durch das IWT in der Datenbank WoS erhoben.

Die Erhebung der Daten erfolgte nach dem „work done at“-Prinzip, d.h. die Forschungsleistung einer Person wurde derjenigen Einrichtung zugerechnet, an der sie erbracht wurde. Entscheidend war somit der angegebene Beschäftigungszeitraum der gemeldeten Wissenschaftler. Wechselte ein Wissenschaftler im Erhebungszeitraum an eine andere Einrichtung, wurde seine nach diesem Wechsel erbrachte Leis-

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

[11] Im laufenden Verfahren zogen insgesamt sechs Forschungseinheiten und eine Universität nachträglich aus verschiedenen Gründen ihre Teilnahme an der Pilotstudie zurück und sind in den genannten Zahlen nicht mehr enthalten.

tung der neuen Einrichtung zugerechnet, jede vorher erbrachte Leistung verblieb bei der ersten Einrichtung. Die Alternative wäre eine Erhebung nach dem „current potential“-Prinzip gewesen. Hierbei wird einer Einrichtung alles angerechnet, was die zu einem bestimmten Stichtag von ihr gemeldeten Wissenschaftler im vorangegangenen Erhebungszeitraum erbracht haben, unabhängig davon, wo sie gerade tätig waren, als sie die berichteten Forschungen durchführten. Ein Wissenschaftler nimmt also bei einem Wechsel gewissermaßen seine Leistungen mit.

Die Geschäftsstelle des Wissenschaftsrates hat die Fachkoordinatoren während der Datenerhebung telefonisch und postalisch betreut und auf ihrer Webseite eine ständig weiterentwickelte Liste häufig gestellter Fragen und Antworten eingerichtet (FAQs).[12] Etwa 60% der teilnehmenden Einrichtungen wandten sich in der zweiten Erhebungsphase mit Fragen an die Geschäftsstelle. Trotz intensiver Betreuung und Begleitung der Erhebung waren bei 66 von 77 Einrichtungen (86%) Korrekturen und Nacherhebungen von Daten erforderlich. Zu den häufigsten Korrekturgründen zählten:

- Angaben außerhalb des Erhebungszeitraums;
- fehlende Angaben von Jahreszahlen / Zeiträumen;
- Unstimmigkeiten zwischen Fragebogen I (Erfassung der leitenden Wissenschaftler) und Fragebogen II (Personalzahlen);
- Unstimmigkeiten zwischen Fragebogen II (summarische Angaben der Einrichtung) und Fragebogen III (Detailinformationen der Forschungseinheiten);
- Unstimmigkeiten zwischen eigenen Angaben und externen Daten;
- nicht gefragte Angaben zur akademischen Selbstverwaltung oder zu von der Einrichtung selbst vergebenen Preisen;
- lückenhafte oder überlange Selbstbeschreibungen;
- lückenhafte Detaildaten, z. B. fehlende Bewilligungssummen oder Projekttitel.

Die Geschäftsstelle teilte den jeweiligen Fachkoordinatoren alle Korrekturen mit einer kurzen Begründung mit und bat, soweit notwendig, um Ergänzungen.

III.4. Publikationserhebung und Zitationsanalyse

Parallel zur Datenerhebung bei den Einrichtungen führte das IWT die Publikationserhebung und anschließend die Zitationsanalyse durch. Auch hier war das leitende Prinzip die „work done at“-Analyse für den Erhebungszeitraum 01.01.2001 bis

[12] Die „Frequently Asked Questions“ sind einzusehen unter: www.wissenschaftsrat.de/pilot_start.htm.

31.12.2005. Den einzelnen Einrichtungen und Forschungseinheiten wurden also jeweils diejenigen Publikationen angerechnet, die die von ihnen gemeldeten leitenden Wissenschaftler während des angegebenen Beschäftigungszeitraums unter der Adresse der Einrichtung publiziert hatten. Dabei spielte es keine Rolle, welchem Fachgebiet die Zeitschrift, in der eine Publikation erschien, zugeordnet war. Die so erstellten Publikationslisten wurden durch die Einrichtungen kontrolliert und ggf. korrigiert. Im Anschluss wurden die Zitationen, die diese Publikationen erhalten hatten, ermittelt. Das Zitationsfenster bezog Zitationen bis Ende des Jahres 2006 ein. Bei der Auswertung der bibliometrischen Daten galten folgende Regeln:

- Kopublikationen von Autoren mehrerer Einrichtungen wurden jeder beteiligten Einrichtung voll angerechnet („normal counting“);
- Kopublikationen, deren Autoren unterschiedlichen Forschungs*einheiten* derselben Einrichtung angehören, wurden diesen dagegen anteilig, zu gleichen Teilen, zugerechnet („fractional counting“);
- das Prinzip der Fraktionierung galt auch für Zitationen von Kopublikationen innerhalb einer Einrichtung;
- Zitationszahlen wurden um Selbstzitationen bereinigt.[13]

Bevor die Publikations- und Zitationsanalyse durchgeführt wurde, wurde unter Abwägung der Vor- und Nachteile etablierter Indikatoren ein möglichst ausgewogenes bibliometrisches Indikatorenset bestimmt. Die nachstehende Tabelle zeigt, welche Indikatoren aus welchem Grund zum Einsatz kamen und führt mögliche Vorbehalte gegenüber einzelnen Indikatoren an. Diese Vorbehalte sowie die Hintergründe der einzelnen Indikatoren wurden den Gutachtern ausführlich erläutert; in Einzelfällen wurden die Basisdaten, die den abgeleiteten Indikatoren zugrunde liegen, hinzugezogen. Diese Transparenz erlaubte es den Gutachtern, die Indikatoren adäquat einzuordnen und zu bewerten und erhöhte zugleich das Bewusstsein dafür, die bibliometrischen Indikatoren nicht zum alleinigen Gradmesser der Forschungsleistung zu machen.

[13] Ausnahme sind die normierten relativen Zitationszahlen, da die vom WoS bezogenen Normierungsfaktoren für den internationalen Durchschnitt ebenfalls nicht um Selbstzitationen bereinigt sind.

Tabelle 2: Verwendete bibliometrische Indikatoren

Kriterium	Indikator	Grund	Vorbehalte
Forschungsqualität (Ebene Forschungseinheit)	Zitationen pro Publikation, fachnormiert und nach Publikationstypus normiert (ZP/FCS_m)	Zitationserfolg größenunabhängig, auf internationale Community relativiert	Abgrenzung Subfields durch Thomson Scientific
	Zitationen pro Publikation, journalnormiert und nach Publikationstypus normiert (ZP/JCS_m)	Zitationserfolg größenunabhängig, relativ zum gewählten Journal	möglicher unerwünschter Anreiz, in niedrigzitierten Zeitschriften zu publizieren
	$JCSm/FCS_m$	abgeleiteter Indikator für die Publikationsstrategie	
	Hintergrundinformationen: P, P_{noncit}, Z_{max}	Indizien für Breite / Kontinuität des Zitationserfolgs	
Impact/Effektivität[1] **(Ebene Einrichtung)**	Zitationszahl (Z)	Maß für die absolute Ausstrahlung in die Scientific Community	nicht fachnormiert
	Publikationszahl	Maß für die Produktivität	möglicher unerwünschter Anreiz, kleinteilig zu publizieren
	ZP/FCS_m	fachnormierter relativer Impact bezogen auf die Einrichtung	kein Maß für absolute Ausstrahlung, nur als Hintergrundinformation verwendet
Effizienz (Ebene Einrichtung)	Z/VZÄ	Ausstrahlung relativ zum Personaleinsatz	nicht fachnormiert
	P/VZÄ	Maß für die Produktivität relativ zum Personaleinsatz	wie bei absoluter Publikationszahl

1) Impact/Effektivität = Sichtbarkeit der Forschung in der scientific community.

Insbesondere der auf Ebene der Forschungseinheiten vorgenommenen Bewertung der Forschungsqualität lagen nicht nur absolute Werte (Publikationszahlen, Zitationszahlen), sondern auch die relative Zitationsrate (Zitationen pro Publikation) und normierte Werte (Zitationen pro Publikation bezogen auf den Fachgebietsdurchschnitt bzw. auf den Zeitschriftendurchschnitt) zugrunde. Die normierten Zitationsraten erlauben einen Abgleich des nationalen Publikations- bzw. Zitationserfolgs mit dem internationalen Durchschnitt.[14] Bei der Normierung werden die Publikationen einer Forschungseinheit nicht alle dem selben Subfield zugeordnet. Vielmehr wird zunächst jede einzelne Publikation über die Zeitschrift, in der sie erschienen ist, ei-

[14] Vgl. dazu entsprechende Auswertungen in der Anlage: „Forschungsleistungen deutscher Universitäten und außeruniversitärer Einrichtungen in der Chemie. Ergebnisse der Pilotstudie Forschungsrating“.

nem Subfield zugeordnet und erhält einen für dieses Subfield, den jeweiligen Publikationstyp (Article, Review, Letter) und Jahrgang spezifischen Normierungsfaktor zugewiesen. Durch die Aggregation der Normierungsfaktoren der einzelnen Publikationen ergibt sich für jede Forschungseinheit ein individueller Normierungsfaktor, in dem sich ihre disziplinäre Spezialisierung und ihre Publikationsstrategie widerspiegeln. Die fachgebietsnormierten Indikatoren wurden unter Rückgriff auf die vom Datenbankanbieter vorgenommene Zuordnung von Publikationen bzw. Zeitschriften zu bestimmten Subfields berechnet. Diese Subfields werden zwar regelmäßig neu definiert, sind aber stellenweise zu wenig differenziert. Es gibt beispielsweise kein eigenes Subfield „Technische Chemie“. Publikationen in diesem Bereich werden offenbar dem Subfield „Chemical Engineering“ zugeordnet, sind somit also nicht vom Chemieingenieurwesen unterschieden, das aber in der Pilotstudie keine Rolle spielte, da es in Deutschland üblicherweise nicht der Chemie zugeordnet ist. Wichtig war deshalb, dass in die Datenberichte der Einrichtungen (s. u.) zwar nur die festgelegten Indikatoren eingingen, in Einzelfällen aber auch Zusatzinformationen wie die Zitationswerte der einzelnen Publikationen einer Forschungseinheit zur Bewertung herangezogen wurden, etwa um auszuschließen, dass bei einer insgesamt nur geringen Publikationszahl die Daten durch „Ausreißer“ oder durch nicht nachvollziehbare Subfield-Zuordnungen verzerrt sind.[15]

Insgesamt wurden in der Publikationsanalyse 41.948 Publikationen und 320.722 Zitationen (ohne Selbstzitationen) erfasst; damit ist die in der Pilotstudie durchgeführte bibliometrische Auswertung auf dem Gebiet der deutschen Chemie die bisher ausführlichste.

III.5. Aufbereitung der Daten

Im Anschluss an die umfangreiche Datenerhebung bereitete die Geschäftsstelle des Wissenschaftsrates die Daten auf. Die bei den Einrichtungen erhobenen und die externen Daten wurden partiell zu abgeleiten Indikatoren verrechnet, alle Angaben den Bewertungsebenen (Einrichtung versus Forschungseinheit) zugeordnet und zu sogenannten „Datenberichten“ jeder einzelnen Einrichtung zusammengefasst, die sich in die sechs Bewertungskriterien gliederten. Zusätzlich enthielten die Datenberichte die Rahmeninformationen der Einrichtung und ihrer Forschungseinheiten. Im Mittel waren die Datenberichte je Einrichtung etwa 50 (20-120) Seiten lang, wobei die Län-

[15] Nähere Informationen zu den bibliometrischen Indikatoren und ihrer Berechnung in den Informationen zur Datengrundlage unter: www.wissenschaftsrat.de/pilot_start.htm.

ge zu einem erheblich Teil auf die Zahl der angegebenen eingeladenen Vorträge zurückzuführen war (s. dazu C.II, S. 59 f.).

Nach Eingang und Kontrolle aller Daten wurden für die quantitativen Indikatoren statistische Lagemaße (Perzentile) ermittelt und in die Datenberichte eingefügt. Weitere Verteilungsmaße (Median, 1. und 3. Quartil) wurden in einem „Leitfaden zu den Datenberichten“ ausgewiesen, so dass sie in der Bewertung berücksichtigt werden konnten.[16]

Die Datenberichte wurden den Einrichtungen zur abschließenden Kontrolle vorgelegt und erst nach erfolgter Rückkoppelung an die Bewertungsgruppe weitergeleitet. Da allerdings bei der Frist der Datenerhebung eine große Kulanz gewährt worden war, musste die Rückkopplungsschleife auf maximal drei, in Einzelfällen auf eine Woche verkürzt werden.

A.IV. Bewertungsphase

IV.1. Organisation und Ablauf

Die Bewertung der Forschungsleistungen auf Basis der Datenberichte war Aufgabe der Bewertungsgruppe. Die Bewertungen wurden arbeitsteilig vorgenommen: Jeder Einrichtung und jeder Forschungseinheit wurden mindestens zwei Berichterstatter aus der Bewertungsgruppe zugeordnet, ausgehend von deren Expertise und orientiert an den Teilgebietszuordnungen der Forschungseinheiten. Dabei wurden Befangenheiten ausgeschlossen. Auch wurden keine außeruniversitären Einrichtungen von Angehörigen derselben Trägerorganisation bewertet. Wenn es aufgrund von Befangenheiten oder fehlender Expertise nicht möglich war, Gutachter aus der Bewertungsgruppe zuzuordnen, wurde auf externe Sondergutachter zurückgegriffen. Auf Sondergutachter wurde auch dann zurückgegriffen, wenn ein Dissens der Berichterstatter nicht aufzulösen war. Insgesamt wurden für das Fach Chemie neun Sondergutachter hinzugezogen.

Die Bewertung basierte nicht nur auf den Datenberichten jeder Einrichtung und den im Leitfaden angegebenen Verteilungsmaßen, sondern auch auf den Publikationslisten jeder Einrichtung bzw. Forschungseinheit, die von den Gutachtern gesichtet wurden. Das IWT stellte in Einzelfällen, bei denen die Interpretation der bibliometrischen Indikatoren problematisch schien, zusätzlich die Basisdaten der Zitationsanalyse zur

[16] Der „Leitfaden zu den Datenberichten für die Bewertungsgruppe Chemie“ ist einzusehen in den Informationen zur Datengrundlage unter http://www.wissenschaftsrat.de/pilot_start.htm.

Verfügung. In Fällen, in denen eine Unsicherheit bestand, wurden darüber hinaus Publikationen zur Bewertung herangezogen.

Die Bewertung war in zwei Phasen eingeteilt: In einer ersten individuellen Bewertungsphase bewerteten die Berichterstatter unabhängig voneinander die ihnen zugeordneten Forschungseinheiten und Einrichtungen über ein durch das Zentrum für Evaluation und Methoden (ZEM) der Universität Bonn erstelltes Online-Eingabeformular bzw. über ein Papierformular. Die Bewertung erfolgte auf der 5-stufigen Notenskala, wobei in der individuellen Benotung noch Zwischennoten vergeben werden konnten (etwa 3,5 = „gut bis sehr gut“), die in der zweiten Bewertungsphase auf ganze Noten bereinigt wurden. Auch wurden in der individuellen Bewertungsphase noch Noten für jeden der in der Bewertungsmatrix verzeichneten Bewertungsaspekte (=Teilaspekte eines Kriteriums) vergeben, wobei es keine Vorgabe gab, mit welcher Gewichtung die Beurteilung eines Einzelaspekts in die Gesamtnote eines Kriteriums eingehen sollte. Bei der Beurteilung der quantitativen Daten wurden keine Vorgaben hinsichtlich der Notenverteilung gemacht, es gab also bspw. keine Vorfestlegung, dass etwa die Top 5% als „exzellent“ eingestuft werden mussten.

In der zweiten Bewertungsphase stimmte die gesamte Bewertungsgruppe über alle individuellen Bewertungen plenar ab. Dafür wurden insgesamt drei je zwei-tägige Sitzungen abgehalten. In der dritten dieser Sitzungen konsolidierte die Bewertungsgruppe die zuvor nur vorläufig festgehaltenen Extremnoten „exzellent“ und „nicht befriedigend“ im direkten Vergleich. Schließlich führte sie einen abschließenden Abgleich aller vorgenommenen Bewertungen durch. Dadurch wurde jede Bewertung zweimal besprochen, d. h. alle 349 Bewertungen nach Kriterium I (Ebene Forschungseinheit) und die jeweils 77 Bewertungen nach den Kriterien II-VI wurden zweimal aufgerufen. Die Berichterstatter begründeten jeden Notenvorschlag kurz, bevor das Plenum über die endgültige Note diskutierte und abstimmte. Befangene Gutachter nahmen nicht an den Beratungen über die jeweiligen Einrichtungen teil. Erwartungsgemäß war der Diskussionsbedarf bei hohem Gutachter-Dissens in der Vorbenotung höher als bei übereinstimmenden Bewertungen. Die Übereinstimmung der individuellen Gutachterurteile in der Bewertung war insgesamt hoch: Nur in ca. 14-26% der Fälle (je nach Kriterium) waren Abweichungen von einer Note und mehr zu verzeichnen, d.h. in ca. 74-86% der Bewertungen stimmten die Berichterstatter bis auf eine maximale Differenz von einer halben Note überein.

Tabelle 3: Übereinstimmung der Gutachterurteile

Kriterium	Anzahl übereinstimmend bewerteter Einr. / FE [1)]	Prozentsatz	Anzahl übereinstimmend bewerteter Einr. / FE ohne "nicht bewertbar" [1) 2)]	Prozentsatz ohne „nicht bewertbar“ [2)]
Kriterium I Forschungsqualität	281 FE	80,5%	292 FE	86,4%
Kriterium II Impact/Effektivität[3)]	63 Einr.	81,8%	63 Einr.	81,8%
Kriterium III Effizienz	56 Einr.	72,7%	57 Einr.	75,0%
Kriterium IV Nachwuchsförderung	57 Einr.	74,0%	60 Einr.	81,1%
Kriterium V Transfer	54 Einr.	70,1%	56 Einr.	74,7%
Kriterium VI Wissensvermittlung	nicht anwendbar wegen der in der plenaren Bewertungsphase vorgenommenen Änderung der Skala			

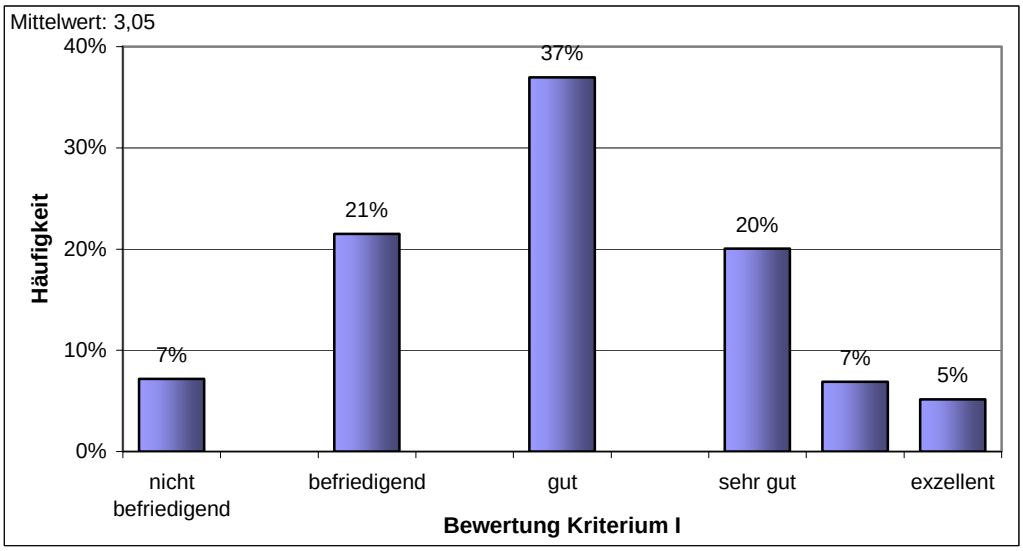
1) In der individuellen Bewertung konnten auch Zwischennoten vergeben werden; Abweichungen innerhalb einer Notenstufe werden hier nicht als abweichende Urteile gewertet, sondern gelten als Übereinstimmung.
2) Das Urteil „nicht bewertbar“ wird nicht als abweichendes Urteil betrachtet, da es keine Wertung enthält.
3) „Impact/Effektivität“ = Sichtbarkeit der Forschung in der scientific community.

Bei der Bewertung von Forschungsqualität und Impact/Effektivität ist die Übereinstimmung der jeweils zwei Berichterstatter am größten. Die niedrigste Übereinstimmung gibt es beim Kriterium Transfer in andere gesellschaftliche Bereiche. Insgesamt beträgt die Gutachter-Übereinstimmung ca. 80%.

IV.2. Analyse der Bewertungsergebnisse

Die durch das Plenum abschließend abgestimmten Bewertungsergebnisse streuen je nach Bewertungskriterium unterschiedlich.

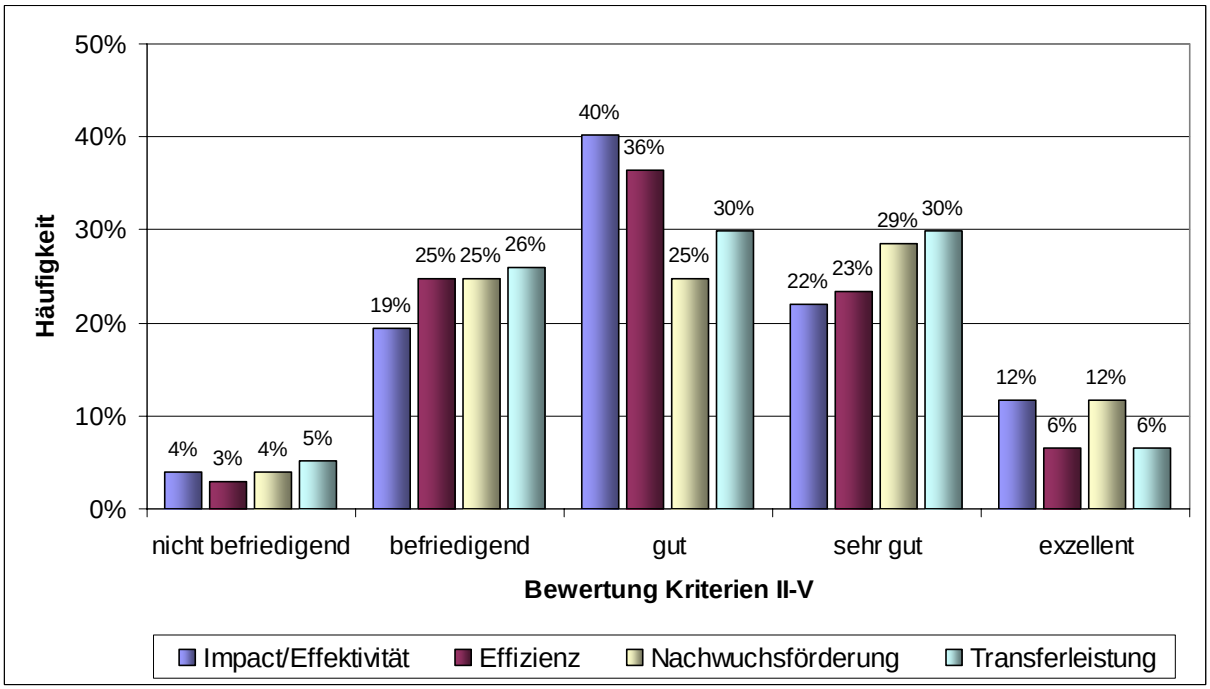
Abbildung 1: Verteilung der Bewertungen zu Kriterium I Forschungsqualität



Die Verteilung der Bewertungen für Kriterium I Forschungsqualität ist, basierend auf 349 Fällen (Forschungseinheiten), breit gestreut, d.h. das Notenspektrum wurde voll ausgeschöpft.

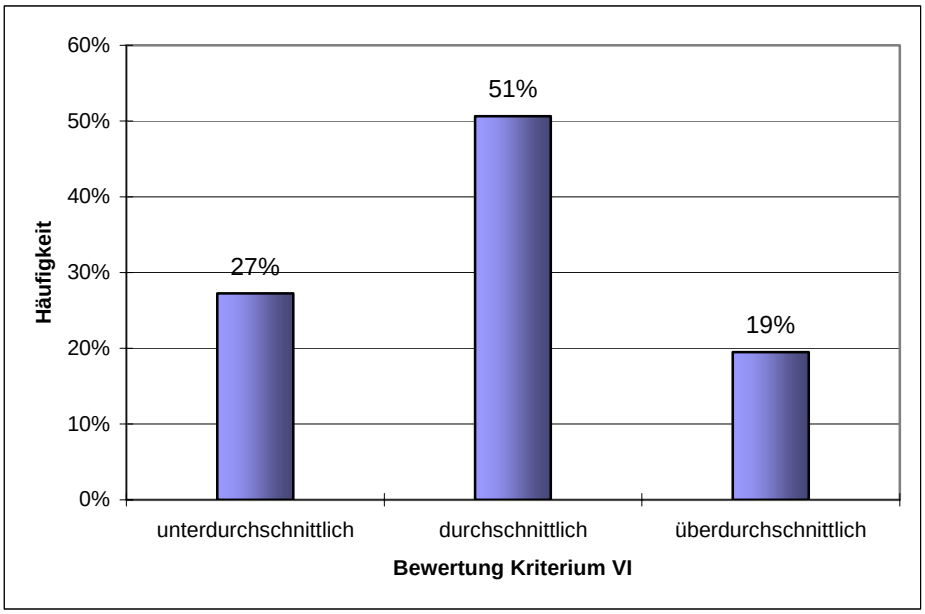
Die nachfolgende Übersicht zeigt für die bzgl. der Notenskala vergleichbaren Kriterien II bis V die jeweiligen Notenverteilungen auf Einrichtungsebene im Vergleich:

Abbildung 2: Verteilung der Bewertungen zu Kriterien II bis V im Vergleich



Auffallend ist vor allem, dass bei den Kriterien II Impact/Effektivität und III Effizienz die mittlere Note „gut“ deutlich am häufigsten vergeben wurde, während die Bewertungen für Kriterium IV Nachwuchsförderung und für Kriterium V Transferleistung etwa zu gleichen Teilen auf die Notenstufen „befriedigend“, „gut“ und „sehr gut“ entfallen. Die Bewertungen für Kriterium VI Wissensvermittlung, die auf drei Notenstufen reduziert wurden, verteilen sich folgendermaßen:

Abbildung 3: Verteilung der Bewertungen zu Kriterium VI Wissenstransfer



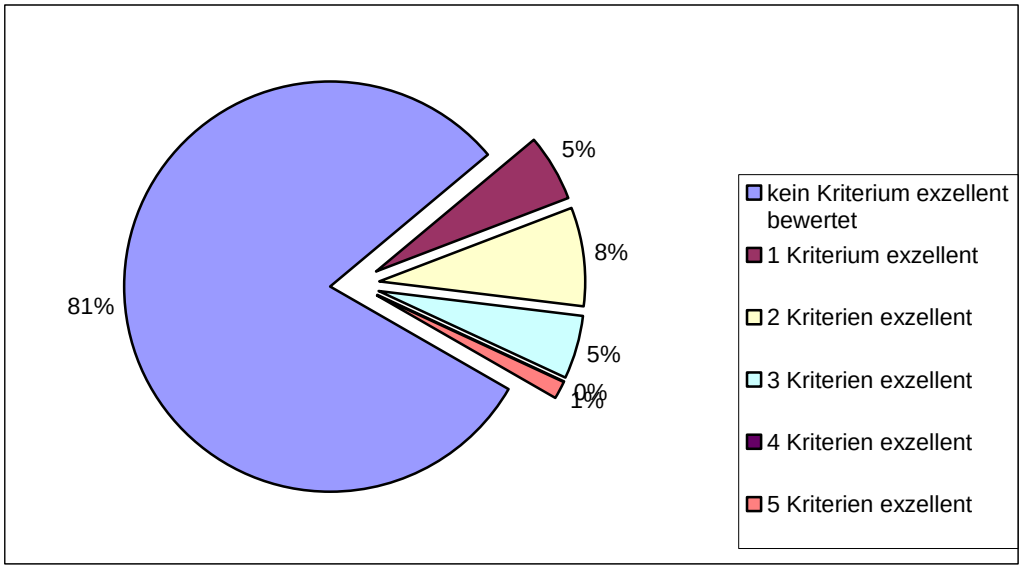
Insgesamt liegen die hier ausgewiesenen Kriterien im Mittel leicht oberhalb der mittleren Note „gut“ (=3). Die durchschnittlich höchsten Bewertungen wurden in den Kriterien Impact/Effektivität (3,19) und Nachwuchsförderung (3,21) gegeben. Die Unterschiede der Mittelwerte der einzelnen Kriterien sind statistisch nicht signifikant.[17]

Je nach Kriterium wurden 5-12% der Einrichtungen bzw. Forschungseinheiten mit „exzellent“ bewertet. Insgesamt erreichten 15 Einrichtungen in mindestens einem Kriterium die Bewertung „exzellent“.[18] Nur eine Einrichtung schaffte es, in allen fünf Kriterien (Kriterium VI ausgenommen) mit „exzellent“ bewertet zu werden. Das nachfolgende Diagramm verdeutlicht, wie viele Einrichtungen in keinem oder bis zu fünf Kriterien die Bewertung „exzellent“ erhielten.

[17] Eine einfaktorielle Varianzanalyse hat im Vergleich der Bewertungen der verschiedenen Kriterien keinen Zusammenhang zwischen bewertetem Kriterium und Bewertungsergebnis gezeigt.

[18] Nicht berücksichtigt sind die Einrichtungen, die für mindestens eine Forschungseinheit die Bewertung „exzellent“ erhielten, sofern sich dies nicht in einem gewichteten Mittelwert von über 4,5 (=exzellente Forschungsqualität auf Ebene der Einrichtung) niederschlägt.

Abbildung 4: Häufigkeit der Bewertung „exzellent“

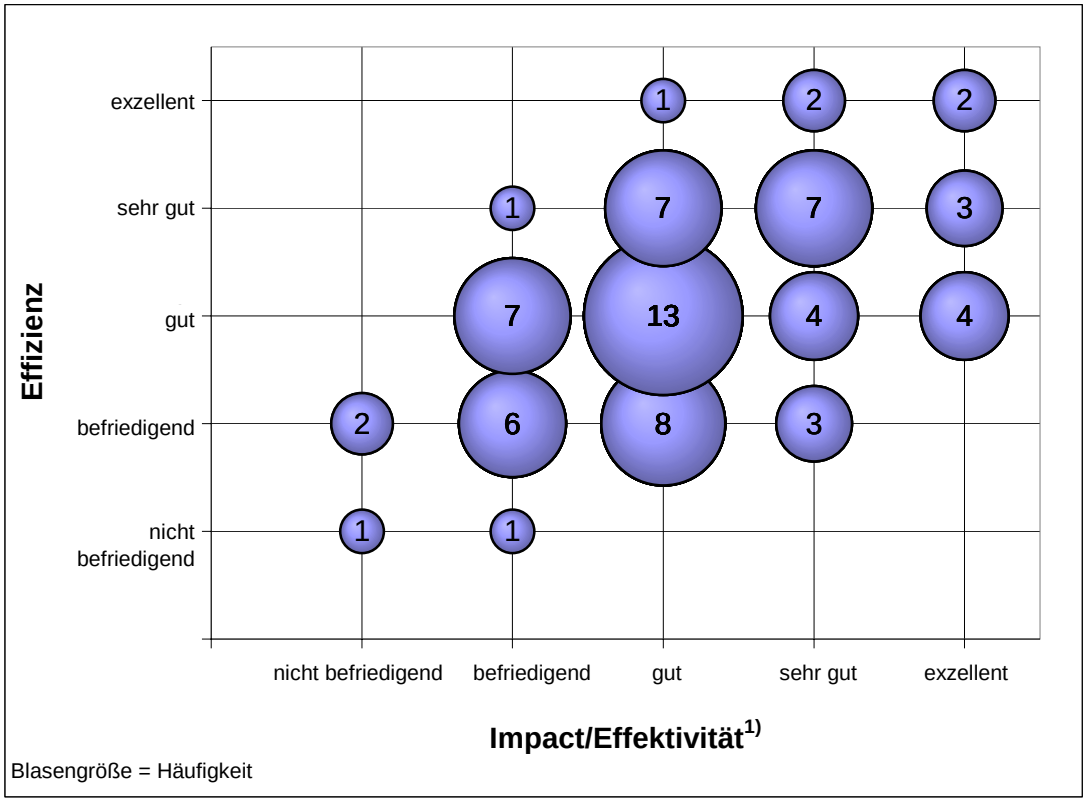


Kriterium „Wissensvermittlung und -verbreitung“ ausgenommen, da darin die Bewertung „exzellent“ nicht vergeben wurde.

Jeweils neun Einrichtungen sind in den Kriterien Impact/Effektivität und Nachwuchsförderung mit „exzellent“ bewertet, vier im Kriterium Forschungsqualität (gewichteter Mittelwert der Forschungseinheiten auf Ebene der Einrichtung), und jeweils fünf in den Kriterien Effizienz und Transfer.

Statistisch lässt sich nachvollziehen, ob es bei den Bewertungen einen dominanten Zusammenhang etwa inhaltlich über eine Dimension miteinander verbundener Kriterien gibt. Nachfolgendes Diagramm zeigt, dass nur bei zwei Einrichtungen sowohl Impact/Effektivität als auch die Effizienz – und somit zwei der Kriterien der Dimension Forschung – mit „exzellent“ bewertet wurden, bei insgesamt neun bzgl. ihres Impacts/ihrer Effektivität als „exzellent“ und fünf bzgl. ihrer Effizienz als „exzellent“ bewerteten Einrichtungen.

Abbildung 5: Korrelation der Bewertungen von Impact/Effektivität und Effizienz



Die in den Blasen angegebenen Werte geben Auskunft über die Häufigkeit einer bestimmten Kombination von Bewertungen.
1) „Impact/Effektivität" = Sichtbarkeit der Forschung in der scientific community.

Bei denjenigen Einrichtungen, deren Impact/Effektivität gleich ihrer Effizienz ist, wurde überwiegend auch die mittlere gewichtete Forschungsqualität ähnlich bewertet wie Impact/Effektivität und Effizienz. Bei den Einrichtungen, deren Impact-/Effektivitätsbewertung von der Effizienzbewertung abweicht, ist die mittlere gewichtete Bewertung der Forschungsqualität eher mit der Bewertung von Impact/Effektivität assoziiert.

Die Ermittlung der Korrelationen der Bewertung der einzelnen Kriterien untereinander bestätigt diesen Befund: Die Korrelation der Bewertungen von Impact/Effektivität und Forschungsqualität (gewichtetes Mittel) ist mit einem Koeffizienten von 0,769 recht hoch. Demgegenüber besteht eine eher geringe Korrelation zwischen der Forschungsqualität (Mittel) und der Effizienz-Bewertung (r=0,498).[19]

Die höhere Korrelation von gewichteter mittlerer Forschungsqualität und Impact/Effektivität lässt den vorsichtigen Schluss zu, dass an Einrichtungen, die insgesamt forschungsstark sind und über eine hohe Ausstrahlung verfügen, in der Regel auch die Forschungsqualität der einzelnen Forschungseinheiten höher ist. Es gibt

[19] Die nach Spearman-Rho ermittelten Korrelationen sind auf dem Niveau von 0,01 2-seitig signifikant.

aber auch schwache Einheiten an insgesamt starken Einrichtungen; umgekehrt verfügen vier Einrichtungen mit einem „guten“ Impact/Effektivität über einzelne Forschungseinheiten, deren Forschungsqualität als „exzellent“ oder „sehr gut bis exzellent“ bewertet wurde. Auch die Nachwuchsförderung korreliert stark mit der Bewertung von Impact/Effektivität. Hingegen ist die Effizienz deutlich geringer mit dem Kriterium Impact/Effektivität korreliert, was bekräftigt, dass es sinnvoll ist, absolute und relative Leistungen unabhängig voneinander zu bewerten. Die Bewertung der Wissensvermittlung hängt insgesamt am wenigsten stark mit den anderen Bewertungen – auch mit dem anderen Kriterium der selben Dimension, der Transferleistung – zusammen, so dass eine Einrichtung trotz schlechter Bewertungen in diesem Kriterium in den anderen Kriterien durchaus sehr gute Bewertungen erlangen kann und umgekehrt.[20]

Die Korrelation der Kriterien untereinander verdeutlicht einerseits deren Zusammenhang, andererseits zeigt sich, dass es möglich war, die Kriterien unabhängig voneinander zu bewerten und dass diese ein ausreichend differenziertes Gesamtprofil generieren. Es gibt keinen dominanten Zusammenhang von Kriterien, der auf eine eindeutige Redundanz hinweisen würde.

In allen sechs Kriterien gab es einzelne Einrichtungen oder Forschungseinheiten, die die Gutachter als „nicht bewertbar“ einstuften. Dafür gab es verschiedenste Gründe:

- Vier Forschungseinheiten wurden nach dem Kriterium I Forschungsqualität nicht bewertet, weil sie erst gegen Ende des Erhebungszeitraumes institutionalisiert bzw. neu besetzt wurden.
- Eine Forschungseinheit wurde aufgrund fehlender Daten nach dem Kriterium I Forschungsqualität nicht bewertet.
- Eine Einrichtung wurde aufgrund fehlender Daten nach dem Kriterium III Effizienz nicht bewertet.
- Eine Einrichtung wurde aufgrund nicht nachvollziehbarer Daten nach dem Kriterium III Effizienz nicht bewertet.
- Eine Einrichtung wurde nach dem Kriterium III Effizienz nicht bewertet, weil die Personalressourcen einer Service-Einheit nicht eindeutig zuzuordnen waren.
- Drei außeruniversitäre Einrichtungen wurden nach dem Kriterium IV Nachwuchsförderung nicht bewertet, da die erhobenen Indikatoren vor allem die Förderung des akademischen Nachwuchses abdecken. Die Förderung des akademischen

[20] Bei der Berechnung der Korrelation ist der Skalen-Unterschied (5- bzw. 3-skalig) irrelevant.

Nachwuchses dieser Einrichtungen wird an den benachbarten Universitäten registriert.

- Zwei Einrichtungen wurden aufgrund fehlender oder fragwürdiger Datengrundlage nach allen Kriterien I-VI nicht als ganze Einrichtungen bewertet. Hier liegen nur Ergebnisse einzelner Forschungseinheiten vor.

A.V. Veröffentlichung und Reaktionen

V.1. Veröffentlichung der Ergebnisse

Die Bewertungsergebnisse wurden am 18. Dezember 2007 durch die Steuerungsgruppe veröffentlicht. Flankiert wurde die Veröffentlichung durch eine zusammenfassende Analyse der Bewertungsergebnisse im Sinne einer Stärken- und Schwächenanalyse der Chemie in Deutschland, die von der Bewertungsgruppe vorbereitet worden war. Der Bericht enthält neben den Ergebnissen des Ratings auch eine kurze Erläuterung des in der Pilotstudie erstmals angewandten Verfahrens. Dieser Bericht wurde am 12. November 2007 durch die Steuerungsgruppe verabschiedet.

Zusätzlich zu den generellen Schlussfolgerungen und Verfahrenserläuterungen wurden die Einzelergebnisse in Form von zwei Diagrammen für jede Einrichtung veröffentlicht, aus denen das individuelle Bewertungsprofil jeder Einrichtung sichtbar wird. Ein erstes Diagramm zeigt die Bewertung der Kriterien I bis VI in einem Balkendiagramm. Neben den Einzelwerten der jeweiligen Einrichtung enthält das Diagramm auch Balken, die die jeweiligen Mittelwerte über alle Einrichtungen darstellen. Dadurch ist ein Abgleich der Leistung der Einrichtung mit der Durchschnittsleistung möglich. Da Kriterium I auf Ebene der Forschungseinheiten bewertet wurde, ist es für die Gesamtdarstellung des Kriteriums Forschungsqualität der Einrichtung notwendig, einen Mittelwert der Forschungsqualität der Forschungseinheiten dieser Einrichtung zu errechnen. Dabei wurde nach der Anzahl der zum Stichtag beschäftigten leitenden Wissenschaftler einer jeweiligen Forschungseinheit gewichtet. Forschungseinheiten, die als „nicht bewertbar“ beurteilt wurden, wurden aus dieser Berechnung ausgeklammert. Ein zweites Diagramm liefert einen genaueren Überblick über die Bewertung der Forschungsqualität: Es wird angezeigt, wie viel Prozent einer Einrichtung welche Bewertung in der Forschungsqualität erhält. Auch hier sind die Einzelbewertungen der Forschungseinheiten zur Ermittlung der prozentualen Verteilung nach der Anzahl der leitenden Wissenschaftler zum Stichtag gewichtet. Die in den Diagrammen an 100 fehlenden Prozent sind die als „nicht bewertbar“ klassifizierten

Forschungseinheiten. In einem dritten Diagramm werden die Einzelbewertungen der Forschungseinheiten dargestellt. So wird nicht nur sichtbar, wie genau welche Forschungseinheit bewertet wurde, sondern auch, wie die Bewertung einer jeweiligen Forschungseinheit im Verhältnis zu den anderen Forschungseinheiten einzuordnen ist. Dieses Diagramm wurde nur den Einrichtungen selbst und den jeweiligen Sitzländern bzw. den Trägerorganisationen FhG, HGF, MPG und WGL zum internen Gebrauch vorgelegt. Es wurde den Einrichtungen aber nahe gelegt, auch diese detaillierte Darstellung der Öffentlichkeit zugänglich zu machen. Die Geschäftsstelle des Wissenschaftsrates bot den Einrichtungen an, diese Veröffentlichung über eine zentrale Linkliste auf den Internetseiten der Geschäftsstelle des Wissenschaftsrates zu vernetzen.

Neben den Diagrammen enthält die Veröffentlichung der Bewertungsergebnisse in Einzelfällen Kommentare zur Bewertung. Darin werden eventuelle Besonderheiten der Bewertung der Einrichtung erläutert. Soweit erforderlich, wurden den Einrichtungen zusätzlich auch auf einzelne Forschungseinheiten bezogene Kommentare vorgelegt, die nicht veröffentlicht wurden.

Die Einrichtungen, die Länder und die Trägerorganisationen erhielten ihre eigenen Bewertungsergebnisse eine Woche vor der Veröffentlichung. Die Einrichtungen selbst erhielten außerdem den jeweiligen „Datenbericht“, der der Bewertung zugrunde lag. Dieser erfasst die in den Fragebögen und aus externen Quellen erhobenen Indikatoren zu den einzelnen Kriterien und gibt für die quantitativen Indikatoren Verteilungswerte an. Die Datenberichte sind vertraulich und werden daher nicht veröffentlicht. Die individuellen Datenberichte können den Einrichtungen selbst aber wichtige Hinweise für ihre weitere Entwicklungsplanung liefern.

Die Veröffentlichung der Ergebnisse fand im Rahmen einer Pressekonferenz statt, an der der Vorsitzende der Steuerungsgruppe, der Vorsitzende der Bewertungsgruppe Chemie und der Generalsekretär des Wissenschaftsrates teilnahmen. An einem anschließenden Pressegespräch nahmen außerdem Vertreter des VCI und der GDCh teil. In den Statements der Teilnehmer und im anschließenden Pressegespräch wurden die wichtigsten Aspekte der Studie deutlich gemacht. GDCh und VCI würdigten die sehr gute Forschungsleistung der deutschen Chemie. Die GDCh forderte die Einrichtungen auf, die Ergebnisse der Pilotstudie im positiven Sinne als Signal zum Aufbruch zu verstehen und Optimierungspotentiale auszuschöpfen. Der VCI betonte das Interesse der deutschen chemischen Industrie, die Forschungsleistungen der Che-

mie auf der Basis einer fundierten und differenzierten Bestandsaufnahme der chemischen Forschung, die das Forschungsrating vorgelegt habe, weiter zu verbessern.

V.2. Reaktionen auf das Forschungsrating

Das Forschungsrating wurde während der Pilotstudie in verschiedenen Foren vorgestellt, darunter der Jahrestagung der GDCh im September 2005, der Chemiedozententagung im März 2006 und dem GDCh-Wissenschaftsforum im September 2007. Nach der Erhebungsphase fand ein erster Erfahrungs-Austausch mit Vertretern des VCI und der GDCh sowie einigen Fachkoordinatoren statt. Am Rande der GDCh-Jahresversammlung wurde gegen Abschluss der Bewertungsphase eine erste vorläufige Bilanz gezogen. Die GDCh richtete außerdem parallel zur Veröffentlichung der Ergebnisse ein Diskussionsforum zur Pilotstudie ein.

Die Fachkoordinatoren wurden gebeten, nach Abschluss der Datenerhebung der Bewertungsgruppe Rückmeldung zum Verfahren zu geben. Zahlreiche Einrichtungen machten von dieser Option Gebrauch und äußerten sich nicht nur zu technischen Fragen bzw. Problemen im Umgang mit den Fragebögen, sondern setzten sich mit dem Verfahren in allgemeiner Hinsicht auseinander. Häufig angesprochene Themen waren der generelle Aufwand und Nutzen solcher Erhebungen (Stichworte: „Evaluitis“, „Rankingmüdigkeit“), das Erhebungsprinzip „work-done-at“, der Umgang mit Fluktuationen sowie die Frage der Gewichtung der einzelnen Kriterien. Außerdem wurde der Ablauf und Zeitplan der Erhebung thematisiert, dabei Kritik am hohen Aufwand geäußert und Vorschläge zur Minderung des Erhebungsaufwandes gemacht. Ein weiterer Kritikpunkt war die Definition der Forschungseinheiten als Erhebungsobjekte, hierbei wurden neben Aufwandsbedenken auch Bedenken hinsichtlich der Einheitlichkeit der Definition geäußert und das Problem der Interdisziplinarität angesprochen. Außerdem gab es Rückmeldungen bzgl. einzelner Daten bzw. Kriterien sowie zur Publikations- und Zitationsanalyse. Die während des Verfahrens gemachten Vorschläge der Fachkoordinatoren sowie der Fachorganisationen gehen in die Teile B und C dieses Berichts ein.

Das Echo auf die Pilotstudie Chemie nach ihrer Veröffentlichung am 18.12.2007 war recht breit und fast durchweg positiv. Es äußerten sich neben der Presse vor allem die teilnehmenden Einrichtungen selbst und die Länder bzw. Trägerorganisationen sowie einige Wissenschaftsorganisationen. Insgesamt 17[21] der teilnehmenden Uni-

[21] Stand: 21. Januar 2008

versitäten und außeruniversitären Einrichtungen veröffentlichten in Pressemitteilungen oder Pressestatements ihre Ergebnisse, allerdings wurden nur in fünf Fällen auch die Ergebnisse der Forschungseinheiten angesprochen, dabei benannte nur die TU München die Ergebnisse aller Forschungseinheiten namentlich, in weiteren Fällen wurden zumindest die gut bis exzellent abschneidenden Forschungseinheiten namentlich benannt. Tendenziell werden in den Pressemitteilungen der Einrichtungen die individuellen Ergebnisse im möglichst positiven Sinne gedeutet. In einigen Fällen wurde angedeutet, künftige Entscheidungen an dem Ergebnis des Ratings zu orientieren.

Weitere Äußerungen kamen von drei Ländern sowie von den Wissenschaftsorganisationen HRK und Junge Akademie. Letztere legte am 18.12.2007 vier Thesen zur Zukunft von Forschungsratings vor, in denen das in der Pilotstudie erprobte Forschungsrating in vielen zentralen Punkten als positives Beispiel angeführt wird. Die Max-Planck-Gesellschaft äußerte sich als einzige Trägerorganisation der teilnehmenden außeruniversitären Einrichtungen in einer Pressemitteilung, in der das gute Ergebnis der Max-Planck-Institute aufgegriffen wurde, dem Verfahren aber nur bedingte Aussagekraft bzgl. der außeruniversitären Forschung beigemessen wird, so dass der hohe Aufwand nicht gerechtfertigt sei.

Im Presseecho (Print und Rundfunk) wird das Forschungsrating häufig in einen Kontext mit anderen Rankingverfahren gestellt und meist positiv diesen gegenüber abgegrenzt. Sofern die Ergebnisse en detail angesprochen werden, ist eine Tendenz erkennbar, die „besten“ bzw. „schlechtesten“ Einrichtungen nennen zu wollen, im Sinne einer Rangliste (auch wenn zuvor darauf verwiesen wird, dass es sich um ein Rating ohne Rangliste handelt).

Zwei Autoren beziehen über die Berichterstattung hinaus dezidiert Stellung zu dem neuen Verfahren. In „DIE ZEIT“ äußert Jan-Martin Wiarda am 19.12.2007: „Schwer zu sagen ist indes, welche der guten Nachrichten schwerer wiegt: dass die deutsche Chemie zur weltweiten Spitze gehört – oder dass der einflussreiche Wissenschaftsrat eine völlig neue Bewertungsmethode entwickelt hat“, der er attestiert, dass schon nach der ersten Veröffentlichung eine Menge für sie spreche. In der „FAZ“ resümiert Heike Schmoll am 19.12.2007: „Der Forschungsvergleich des Wissenschaftsrats ist zwar aufwendig und langwierig, dafür aber seriöser, weil er die außeruniversitäre Forschung miterfasst. Hoffentlich setzt sich das Rating auch für andere Fächer durch.“

B. Empfehlungen

B.I. Generelle Bewertung von Aufwand und Nutzen

Bei der Bewertung des Aufwandes des in der Pilotstudie erprobten Verfahrens ist zu beachten, dass der Pilotcharakter der Studie eine Reihe von experimentellen Komponenten und Absicherungsmaßnahmen notwendig gemacht hat, die bei einer Wiederholung entfallen und somit den Aufwand reduzieren können. Die wesentlichen Träger des Aufwands sind die Gutachter, die bewerteten Einrichtungen durch die Datenerhebung, die Geschäftsstelle und die Bibliometrie. Im Folgenden wird kurz angegeben, welche Faktoren den Aufwand an diesen Stellen determinieren.

Gutachter

Die 15 Gutachter der Bewertungsgruppe Chemie sind während der Pilotstudie im Zeitraum Februar 2006 bis Januar 2008 zu insgesamt acht Sitzungen (elf Sitzungstage) zusammengetroffen. Vier davon umfassten die Operationalisierungsphase einschließlich der Auswertung des Pretests, drei jeweils zweitägige Sitzungen die Bewertungsphase. Die letzte Sitzung diente der Abstimmung des vorliegenden Abschlussberichts. Zu den Sitzungstagen kommt die individuelle Vorbereitung hinzu, die vor allem in der Bewertungsphase einen erhöhten Aufwand bedeutete und allein für diese von den Gutachtern auf etwa 7 Arbeitstage pro Person geschätzt wird. Insgesamt betrug der Arbeitsaufwand 20 bis 25 Tage für jeden Gutachter.

Der Arbeitsaufwand der Gutachter ist partiell dem Pilotcharakter der Studie geschuldet. So mussten zunächst grundlegende Operationalisierungsfragen geklärt, Fragebögen entwickelt und in einem Pretest erprobt sowie die Bewertungsskalen definiert werden. Dies war zeitlich mit einer Dauer von neun Monaten die aufwendigste Phase. Bei einer Wiederholung des Verfahrens könnte durch eine von den Erfahrungen der Pilotstudie ausgehende Standardisierung der Aufwand der Operationalisierungsphase reduziert werden. Ganz verzichtbar ist sie allerdings nicht, da die Kriterien und verwendeten Indikatoren bei einem mehrjährigen Turnus auch dann aktualisiert werden müssten, wenn dasselbe Fach erneut bewertet würde. Zudem half die Operationalisierungsphase den Gutachtern auch, sich mit dem Verfahren vertraut zu machen.

Der Aufwand der Bewertungsphase hängt vor allem von der Zahl der Forschungseinheiten, der Zahl der Kriterien und der Qualität der Daten ab. In der Pilotstudie (349

Forschungseinheiten, sechs Kriterien) hatte jeder Gutachter als Berichterstatter die Verantwortung für etwa 45 Forschungseinheiten und etwa zehn Einrichtungen, die nach fünf Kriterien zu bewerten waren, musste also insgesamt knapp 100 Einzelbewertungen vornehmen. Neben inhaltlichen Gründen (s.u.) ist dies ein weiteres Argument dafür, die Differenziertheit der Bewertung zu begrenzen. Des Weiteren ließe sich der Leseaufwand für die Gutachter auch dadurch reduzieren, dass Informationen, die nach den Erfahrungen aus der Pilotstudie keinen Einfluss auf die Bewertung haben, nicht mehr erhoben werden. Schließlich gab es einzelne Fragen, bei denen die Antworten der Einrichtungen von so heterogener Qualität waren, dass die Gutachter sich zu einer unabhängigen Überprüfung der Angaben gezwungen sahen. Eine Verbesserung der Datenqualität oder der Verzicht auf Daten, die nicht verlässlich zu erheben sind, würde diesen Aufwand reduzieren.

Die Erstellung des vorliegenden Abschlussberichts ist spezifisch für eine Pilotstudie. Bei einer Wiederholung des Verfahrens könnten Erfahrungen der Gutachter aus dem Bewertungsprozess auch ohne umfassenden schriftlichen Bericht in die laufende Verfahrensverbesserung einfließen, so dass eine gesonderte Sitzung entbehrlich wäre.

Bei einer Wiederholung des Verfahrens könnte der Aufwand für die Gutachter auf etwa 15 bis 20 Arbeitstage reduziert werden.

<u>Bewertete Einrichtungen / Datenerhebung</u>

Für das „Informed Peer Review“-Verfahren ist eine aussagekräftige Datengrundlage essentiell, die großenteils bei den betroffenen Einrichtungen erhoben wurde. Die zuständigen Fachkoordinatoren haben den Aufwand der Datenerhebung unterschiedlich hoch eingeschätzt; er rangierte nach ihren Angaben von 60 über mehrere hundert Professoren- und Mitarbeiterstunden bis hin zu 2 bis 3 Mannmonaten für eine ganze Einrichtung. Großen Einfluss auf den Aufwand hatten neben der Größe einer Einrichtung auch die Qualität des internen Controllings und die Unterstützung der Wissenschaftler seitens der Verwaltung.

Zu den wichtigsten Faktoren, die den Aufwand bei den Einrichtungen determinierten, gehört, dass der Inhalt der Befragung nicht vorab bekannt war, so dass ein Teil der Daten ex post rekonstruiert werden musste; die insgesamt zu kurze Vorlaufzeit; sowie technische Schwierigkeiten bei der Bedienung des Erhebungsinstrumentes. Alle diese Aspekte sind offenkundig dem Pilotcharakter der Studie geschuldet. Eine län-

gere Vorlaufzeit sowie die Entwicklung eines komfortableren Erhebungsinstrumentes sollten darum bei einer etwaigen Wiederholung des Verfahrens eingeplant werden.

Ein weiterer zentraler Faktor ist der Inhalt der Datenerhebung. Einige Daten haben sich in der Pilotstudie als wenig belastbar erwiesen und sollten nicht wieder erhoben werden. Vorschläge dazu werden in Teil C gemacht.

Folgenreich war ferner, dass Daten nicht nur für eine Einrichtung insgesamt, sondern auch für einzelne Forschungseinheiten erhoben wurden. Detaillierte Angaben auf dieser niedrig aggregierten Stufe, die häufig erst für das Forschungsrating definiert wurde, lagen bei einigen Einrichtungen nicht vor und mussten aus Angaben einzelner Professoren neu zusammengestellt werden. Dies stellte besonders für Einrichtungen mit einer hohen Fluktuation eine Herausforderung dar. Auch dieses Problem ist mit einer längeren Vorlaufzeit zu lösen, wenn die Einrichtungen kontinuierlich die wichtigsten Daten zentral sammeln. Eine weitere Vereinfachung ist zu erzielen, wenn die Größe der Forschungseinheiten stärker vereinheitlicht wird (vgl. B.III.2, S. 41 ff.) und so die Notwendigkeit, Daten über Kleinsteinheiten zu erheben, entfällt.

Arbeitsintensiv war auch das Abfassen der als Selbstbeschreibungen erbetenen kurzen Textpassagen, die in die Bewertung eingingen oder den Berichterstattern als Hintergrundinformationen dienten. Nur in wenigen Fällen wurden solche Texte aus bereits vorhandenen Publikationen (etwa Forschungsberichten) übernommen. Die Texte bzw. Selbstberichte als Rahmeninformationen waren aber sehr wichtig, da sie den Einrichtungen die Möglichkeit gaben, auf Besonderheiten hinzuweisen, die die Bewertung beeinflussen konnten. Diese Möglichkeit ist für die Akzeptanz des Verfahrens essentiell, weil sie verhindert, dass die Einrichtungen nur nach standardisierten Kennziffern bewertet werden.

Ein kritisch gegen die Pilotstudie insgesamt vorgebrachtes Argument der Fachkoordinatoren war eine gewisse „Ratingmüdigkeit“, die die Bereitschaft der betroffenen Wissenschaftler, an der Datenerhebung mitzuwirken, gemindert habe. In Anbetracht zahlreicher bereits vorhandener und durchgeführter Evaluationen sei fraglich, ob sich der Aufwand für ein weiteres Verfahren überhaupt rechtfertigen lasse. Nach Abschluss der Erhebung wurde indes von einigen Einrichtungen erklärt, dass die aufwendig erhobenen Daten ein wichtiges Instrument der Selbsterkenntnis seien. Die 77 teilnehmenden Einrichtungen verfügten offenbar in sehr unterschiedlichem Maße bereits über zentrale, kontinuierlich aktualisierte Datenbestände zu ihren For-

schungsleistungen. Es wäre wünschenswert, hier wenigstens einen Mindeststandard herzustellen, ohne dass dies mit übermäßigem und unangemessenem Ausbau des Verwaltungsapparats einhergehen muss.

Geschäftsstelle

Die Pilotstudie wurde in der Geschäftsstelle des Wissenschaftsrates unterstützt von einer Projektgruppe mit insgesamt fünf Mitarbeitern (vier VZÄ), die parallel auch die Pilotstudie Soziologie betreute. Neben der Organisation des gesamten Verfahrens wirkte diese Gruppe auch an der Entwicklung von Indikatoren mit und führte die Datenerhebung und -kontrolle durch. Eine ausreichende Geschäftsstellenkapazität ist für die Qualität der Verfahrensentwicklung und -organisation und der Datenerhebung essentiell und trägt auch zur Entlastung der Gutachter bei. Im Rahmen der Pilotstudie war es hilfreich, dass für jedes Fach ein wissenschaftlicher Mitarbeiter als Ansprechpartner zur Verfügung stand, der die Gutachter bei der Lösung fachspezifischer Probleme der Operationalisierung und die Fachkoordinatoren bei der Datenerhebung unterstützen konnte.

Bibliometrie

Die bibliometrischen Daten waren für die Pilotstudie Chemie von zentraler Bedeutung. Der für die Erstellung der Publikationslisten und die anschließende Zitationsanalyse betriebene Aufwand wie auch die Lizenzgebühren für die Nutzung der Datenbank waren voll gerechtfertigt. Angesichts der kontinuierlichen Fortentwicklung der bibliometrischen Indikatoren, an der ein Forschungsrating bei einer Wiederholung unbedingt teilhaben müsste, ist mit einer Verringerung dieses Aufwandes nicht zu rechnen.

Die Kosten für die Pilotstudie Forschungsrating im Fach Chemie sind nur teilweise klar zu beziffern. Die direkten Kosten lagen für beide in der Pilotstudie behandelten Fächer Chemie und Soziologie bei 1,1 Mio Euro einschließlich der bibliometrischen Recherchen und Analysen. Hinzu kommen nicht bezifferbare Kosten für die teilnehmenden Einrichtungen. Deren Rückmeldungen zum Aufwand ergeben kein einheitliches Bild; der Aufwand der Pilotstudie könnte bis zu 2 Mannmonate pro Einrichtung betragen haben. Diese für die Datenlieferung zu berücksichtigenden Kosten sind allerdings eng mit der Erfüllung anderweitiger Berichtspflichten der Einrichtungen verknüpft. Zusätzlich ist der Aufwand für die 15 Gutachter zu veranschlagen, die jeweils 20 bis 25 Arbeitstage investiert haben.

Der diesen Kosten gegenüberstehende direkte Nutzen des Forschungsratings besteht in einer differenzierten, belastbaren Bewertung forschender Einrichtungen mit rund 9.700 Wissenschaftlern (6.800 VZÄ) in Deutschland, die sowohl national als auch international für Wissenschaft und Wissenschaftspolitik die Leistungsfähigkeit der deutschen Chemie transparent macht. Wesentlich für die Belastbarkeit der Ergebnisse ist das gewählte „Informed Peer Review“-Prinzip. Dies bedeutet zwar gegenüber einem rein indikatorenbasierten Ranking einen zusätzlichen Aufwand durch Einbeziehung von Gutachtern. Dem steht aber eine deutlich höhere Verlässlichkeit gegenüber, weil das Verfahren in der Lage ist, Besonderheiten der einzelnen Einrichtungen wie auch Ungenauigkeiten der Datengrundlage zu identifizieren und bei der Bewertung zu berücksichtigen. Zudem ist das Verfahren durch die Mehrdimensionalität in der Lage, große wie kleine Einrichtungen und solche mit unterschiedlicher Mission fair zu vergleichen. Nur auf diese Weise ist auch die vergleichende Bewertung der universitären und der außeruniversitären Forschung möglich. Gerade diese beide Einrichtungstypen umfassende Perspektive ist ein großer Vorteil des Ratingverfahrens, das es erstmals erlaubt, die Leistungsfähigkeit der staatlich geförderten chemischen Forschung in Deutschland insgesamt sichtbar zu machen. Der indirekte, längerfristige Nutzen besteht unter anderem in einer besseren Steuerung der chemisch forschenden Einrichtungen in Deutschland. Die Begutachtung durch renommierte Gutachter ist eher geeignet, Akzeptanz in der Fachgemeinschaft zu finden, als ein rein indikatorenbasiertes Ranking, und wird deshalb auch eine größere Autorität entfalten, wenn es darum geht, schwierige strategische Entscheidungen vorzubereiten.

Dieser Nutzen wird sich nach Ansicht der Bewertungsgruppe Chemie allerdings erst dann voll entfalten, wenn anstelle der jetzt vorliegenden, punktuellen Bewertungen durch nochmalige Bewertungen Trends aufgezeigt und Veränderungen der Forschungsleistung sichtbar gemacht werden können.

B.II. Organisation und Ablauf

Die Organisation, die für die Pilotstudie gewählt wurde, hat sich im wesentlichen bewährt. Größe und Zusammensetzung der Bewertungsgruppe waren für ein Fach von der Größe der Chemie angemessen. Für die Akzeptanz des Verfahrens ist das Renommee der Gutachter und eine gute Abdeckung der Teilgebiete des Fachs zentral. Die Beteiligung der Fachgesellschaften und der großen Wissenschaftsorganisationen an der Auswahl der Gutachter hat sich bewährt und sollte bei einer Wiederholung des Verfahrens beibehalten werden. Es wird davon abgeraten, die Gruppen zu ver-

größern oder den Anteil ausländischer Gutachter zu verringern. Die Anwesenheit eines Beobachters aus der Steuerungsgruppe war für die Einheitlichkeit des Verfahrens wichtig. Die Einrichtung einer gesonderten Unterarbeitsgruppe der Steuerungsgruppe zur Entwicklung der Fragebögen ist dem Pilotcharakter der Studie geschuldet und kann künftig entfallen.

Verbessert werden kann die Organisation der Datenerhebung in manchen der teilnehmenden Einrichtungen. Alle Universitäten und außeruniversitären Institute wurden durch Fachkoordinatoren vertreten, die in der Regel Wissenschaftler aus dem Fach Chemie waren. Dies war deshalb wichtig, weil mit der Gliederung in Forschungseinheiten und der Abfassung von kurzen Selbstbeschreibungen strategische Aufgaben erfüllt werden mussten, die in der Hand der Wissenschaft liegen sollten. Zudem sollten die Wissenschaftler die inhaltliche Letztkontrolle behalten. Gleichzeitig war es aber erkennbar ineffizient, wenn die eigentliche Datenerhebung von den Wissenschaftlern selbst durchgeführt wurde; sie sollte in der Hand der jeweiligen Verwaltung liegen. Unzureichende Unterstützung durch die Verwaltung und ein unterentwickeltes internes Controlling führten dazu, dass die Datenerhebung an einigen Universitäten nur unter hohem persönlichem Einsatz des Fachkoordinators möglich war. In einzelnen Fällen blieben die Daten dennoch lückenhaft. Es wird daher dringend empfohlen, dass sich an der Erhebung künftig sowohl die Wissenschaftler selbst als auch die Verwaltung der Einrichtung beteiligen und miteinander effizient kooperieren. Für die Zukunft sollten die Einrichtungen von vorneherein jeweils zwei Ansprechpartner benennen, je einen aus der Wissenschaft und aus der Verwaltung.

Die Phasen der Pilotstudie waren von unterschiedlicher Dauer. Die erste Phase, die fachspezifische Operationalisierung der Kriterien und die Entwicklung von Fragebögen, dauerte inklusive des Pretests etwa 9 Monate. Dieser Zeitraum kann bei einer Wiederholung im selben Fach deutlich verkürzt werden. Wenn neue Indikatoren entwickelt werden, sollte auf einen Pretest aber nicht verzichtet werden.

Die zweite Phase, die Erhebung der Daten bei den Einrichtungen mit Hilfe der Fachkoordinatoren, wurde in der Pilotstudie in zwei Teilabschnitte aufgetrennt: zunächst wurden die Forschungseinheiten und die ihnen zugeordneten leitenden Wissenschaftler erfasst. Dies erfolgte aus Zeitgründen bereits parallel zum Pretest der später zu verwendenden Fragebögen. Dadurch wurde zwar etwa ein Quartal eingespart, der Nachteil war aber, dass die teilnehmenden Einrichtungen noch nicht genau

wussten, welche Daten sie zu den zu bildenden Forschungseinheiten würden bereitstellen müssen. Eine solche Überlappung sollte deshalb künftig vermieden werden.

Für die eigentliche Datenerhebung wurde den Einrichtungen eine Erhebungsfrist von 14 Wochen gewährt, die in vielen Fällen noch ausgeweitet wurde, um eine möglichst hohe Erfassungsquote zu erhalten. Hier war die Abstimmung mit der Geschäftsstelle des Wissenschaftsrates wichtig. Allerdings führte die großzügige Fristauslegung zu verkürzten Korrekturfristen und bedeutete eine gewisse Benachteiligung der fristgemäß einreichenden Einrichtungen. Problematisch waren auch verspätete Nachmeldungen von Personen, deren Publikationsleistung durch das IWT nicht mehr erhoben werden konnte. Die bibliometrischen Daten konnten aufgrund der verlängerten Erhebungsdauer nicht bereits in die den Einrichtungen zur abschließenden Korrektur vorgelegten Datenberichte eingehen, sondern waren erst in den endgültigen Fassungen für die Berichterstatter enthalten. Künftig ist eine strengere Einhaltung der Erhebungsfristen bei gleichzeitiger Verlängerung der Vorlaufzeit anzustreben, um das gesamte Verfahren zeitlich zu straffen. Die im Rahmen der Pilotstudie geübte Kulanz war indes für die Akzeptanz des Verfahrens wichtig.

Die Bewertungsphase war mit ihrer Einteilung in zwei Phasen (1. individuelle Bewertungsphase, 2. plenare Bewertungssitzungen) der Komplexität eines mehrdimensionalen „Informed Peer Review“-Verfahrens angemessen. Durch die zeitlich recht enge Taktung der plenaren Bewertungsphase, die sich insgesamt über nur 2,5 Monate erstreckte, wurde möglicherweise redundanten Abstimmungsproblemen vorgebeugt, gleichzeitig den Berichterstattern ausreichend Gelegenheit gegeben, sich mit den eigenen Bewertungen und den Bewertungen der anderen Bewertungsgruppenmitglieder auseinanderzusetzen.

B.III. Zu Einzelaspekten des Forschungsratings

III.1. Zur fachspezifischen Operationalisierung des Verfahrens

Der erste Schritt der fachspezifischen Operationalisierung war die Definition des Faches Chemie. Die Einteilung in zehn Teilgebiete machte das Fach vor allem hinsichtlich der Zuordnung von Berichterstattern zu einzelnen Forschungseinheiten handhabbar. Allerdings sind die einzelnen Teilgebiete sehr unterschiedlich groß, entsprechend wurden einigen Teilgebieten sehr viele Forschungseinheiten zugeordnet, anderen nur wenige. Einige Forschungseinheiten aus Spezialgebieten der Chemie, wie etwa der „Geochemie“ oder „Meereschemie“, wurden keinem der Teilgebiete zuge-

ordnet, aber dennoch mit bewertet. In Grenzfällen (bspw. Pharmazeutische/Medizinische Chemie) wurden die Einrichtungen aufgefordert, selbst zu entscheiden, ob die betreffenden Einheiten vorrangig der Chemie zuzuordnen sind. Diese Entscheidung war Wissenschaftlern in der Pilotstudie dadurch erschwert, dass nicht feststand, welche alternative Einordnung möglich gewesen wäre. Einige Universitäten haben zudem Einheiten außerhalb ihrer chemischen Fakultät bzw. ihres chemischen Fachbereichs nicht in die Überlegungen einbezogen. Dies führte insbesondere dazu, dass biochemische Einheiten an biologischen oder medizinischen Fakultäten nicht flächendeckend erfasst wurden. Solche Unsicherheiten bezüglich der Zugehörigkeit zur Chemie könnten deutlich vermindert werden, wenn nicht nur die Chemie durch Angabe von Teilgebieten bestimmt, sondern eine vollständige Taxonomie der Fächer für ein Forschungsrating entwickelt und die Erfassung von benachbarten, im Idealfall von allen Fächern synchronisiert würde.

Allerdings waren nicht nur fachliche Zuordnungsschwierigkeiten der Grund für das Fehlen einiger Forschungseinheiten am Rating: in Einzelfällen waren offenkundig einrichtungsinterne strategische Erwägungen ausschlaggebend für die Entscheidung, bestimmte Teilgebiete nicht am Rating zu beteiligen.

Bei der Bestimmung von Indikatoren, die eine adäquate Bewertung des Faches Chemie erlauben, wurden die Spezifika des Faches berücksichtigt. Die Beurteilung von Forschungsqualität anhand bibliometrischer Daten ist in der Chemie aufgrund ihrer Publikationspraxis gut möglich. Auch die Relevanz des Transfers von Wissen ist greifbar über Indikatoren wie Industriemittel, Patente und Lizenzen. Hier besteht allerdings in vielen Fällen Verbesserungsbedarf hinsichtlich des einrichtungsinternen Controllings. Eine fachspezifische Zuordnung von Indikatoren nach zwei Kriterien, nämlich erstens ihrer Aussagekraft für das jeweilige Fach und zweitens ihrer Erhebbarkeit bei den teilnehmenden Institutionen oder aus externen Datenquellen muss der erste Schritt bei der Anpassung des Verfahrens an alle Fächer sein. Dennoch sollte es einen einheitlichen Rahmen geben, der für verschiedene Fächer gleichermaßen gilt.

Der durchgeführte Pretest war wichtig, um die Bewertungsmatrix für das Fach Chemie zu entwickeln, da damit überprüft werden konnte, ob erstens die Erhebung der Indikatoren und zweitens die Bewertung anhand der Matrix möglich ist. Die darauf erfolgten Änderungen waren zwar nicht grundlegend, aber wichtig vor allem zur Begrenzung des Aufwandes. Außerdem konnten sich die Gutachter im Verlauf

des Pretests mit dem Verfahren vertraut machen und in ihrem Bewertungsverhalten eine gemeinsame Linie entwickeln.

III.2. Zur Erfassung der Forschungseinheiten

Die Erfassung der Forschungseinheiten stellte eines der Hauptprobleme sowohl in der Datenerhebung als auch in der Bewertung der Forschungsleistung dar. Die Problemfelder waren folgende:

- die niedrige Aggregationsebene Forschungseinheit erschwerte stellenweise die Erhebung (Bsp. Drittmittelprojekte);
- der heterogene Zuschnitt der Forschungseinheiten (teils unterkritische Forschungseinheiten gebildet, teils zu viele heterogene Arbeitsgruppen zu einer Forschungseinheit zusammengefasst) erschwerte die einheitliche Bewertung;
- die Zuordnung von interdisziplinär arbeitenden Arbeitsgruppen / Wissenschaftlern zu einer chemisch tätigen Forschungseinheit war teils schwierig;
- der Umgang mit Service-Einheiten, die nicht nur Dienstleistungen für die Chemie erbringen, war in der Bewertung problematisch;
- der Umgang mit institutionenübergreifenden Kooperationen war sowohl in der Erhebung als auch in der Bewertung schwierig;
- die möglichen Konsequenzen des Zuschnitts der Forschungseinheiten für die Bewertung waren nicht allen Einrichtungen vorab klar.

Diese Schwierigkeiten hatten indes nur in Einzelfällen tatsächlich Folgen für den Bewertungsprozess. In den meisten Fällen konnten die Wissenschaftler eindeutig den Forschungseinheiten zugeordnet, die Leistungen der Service-Einheiten adäquat auf die Forschungseinheiten verteilt sowie institutionenübergreifende Forschungseinheiten korrekt definiert und ihre Forschungsleistungen eindeutig angerechnet werden.

Die Zuordnung disziplinär uneindeutiger Forschungseinheiten dürfte leichter werden, sobald eine vollständige Taxonomie aller für eine Bewertung in Frage kommender Fächer vorliegt. Wünschenswert wäre die Option, einzelne interdisziplinäre Forschungseinheiten von Gutachtern zweier Fächer bewerten zu lassen. Dies ließe sich allerdings nur realisieren, wenn Ratings der relevanten Fächer zeitgleich stattfänden.

Im außeruniversitären Bereich bereitete zum Teil die Personalkategorie „Gruppenleiter“ als Untergruppe der namentlich zu erfassenden „leitenden Wissenschaftler“ Schwierigkeiten. Diesem Problem könnte man am besten dadurch begegnen, dass

man die namentliche Erfassung für die Publikationsrecherche von der Erhebung der Personalstruktur trennt und die Einrichtungen auffordert, zusätzlich zu den leitenden Wissenschaftlern alle selbstständig forschenden und publizierenden Wissenschaftler unabhängig von ihrem Status namentlich zu melden.

Um die Schwierigkeiten bei der Definition von Forschungseinheiten zu reduzieren und gleichzeitig Manipulationen durch strategischen Zuschnitt der Forschungseinheiten vorzubeugen, sollten die Forschungseinheiten künftig stärker standardisiert werden. Einzelne Lehrstühle / Professuren sollten grundsätzlich nicht mehr als Forschungseinheiten akzeptiert werden. Es ist allerdings nicht möglich und auch nicht nötig, für jeden Einzelfall Regeln im Voraus zu definieren. Verfahrensmäßig ließe sich die Standardisierung dadurch verbessern, dass die von den Einrichtungen vorgeschlagene Forschungseinheitenstruktur durch die Gutachter überprüft wird, bevor auf ihrer Basis die Datenerhebung beginnt.

Eine solche Vorgehensweise würde zugleich ein zweites Problem, das im Zusammenhang mit der Bewertung von Forschungseinheiten nach dem Kriterium Forschungsqualität aufgetreten ist, beheben: das Problem personenbezogener Bewertung und somit datenschutzrechtliche Bedenken. Etwa 10 % der für die Pilotstudie gebildeten Forschungseinheiten bestanden aus weniger als drei leitenden Wissenschaftlern. In diesen Fällen ist die Bewertung einer Forschungseinheit kaum von der Bewertung der dahinter stehenden leitenden Wissenschaftler zu trennen. Eine Veröffentlichung personenbeziehbarer Bewertungen ist jedoch datenschutzrechtlich bedenklich. Daher wurde für die Pilotstudie Forschungsrating insgesamt darauf verzichtet, die Bewertungen der Forschungseinheiten durch die Steuerungsgruppe zu veröffentlichen. Die Bewertungsgruppe Chemie ist aber von der hohen Aussagekraft der Bewertungen der Forschungseinheiten überzeugt. Gerade bei einem deutlich strukturierten Fach wie der Chemie ist es nicht unerheblich, Bewertungen auf einzelne Teilgebiete des Faches beziehen zu können und so etwa herausragende Leistungen in einem Teilgebiet unter insgesamt schwächeren Teilgebieten einer Einrichtung offenzulegen. Zwar ist diese Differenzierung nicht für alle Adressaten des Ratings gleichermaßen relevant, aber für internationale wie nationale Kooperationspartner, für die Anwerbung ausländischer Wissenschaftler sowie für fachfremde interne Entscheidungsträger ist auch diese Information sicherlich hilfreich. Von den genannten Adressaten werden in der Pilotstudie nur die letztgenannten auch über die individuelle Bewertung der Forschungseinheiten informiert. Es wird den teilnehmenden Einrichtungen selbst überlassen, diese Ergebnisse zu veröffentlichen. Die Bewertungs-

gruppe Chemie würde es begrüßen, wenn durch eine angemessene Definition der Forschungseinheiten mit einer Größe, die keine Personenbeziehbarkeit mehr zulässt, die Veröffentlichung der Binnendifferenzierung der Bewertung des Kriteriums Forschungsqualität möglich würde.

Die Bewertungsgruppe Chemie spricht sich aus diesen Gründen dafür aus, Forschungseinheiten vorzugsweise an bestehenden Organisationseinheiten zu orientieren, keine einzelnen Professuren bzw. Lehrstühle zu benennen und insgesamt keine Forschungseinheiten mit einer Größe von weniger als drei leitenden Wissenschaftlern zu definieren. Damit diese Vorgaben eingehalten werden, sollte die vorgeschlagene Struktur einer Einrichtung mit der Geschäftsstelle oder der Bewertungsgruppe rückgekoppelt werden, bevor auf dieser Basis die eigentliche Datenerhebung beginnt. Dieses Vorgehen würde die Datenerhebung vereinfachen, den Erhebungsaufwand verringern, die Bewertungsgrundlage vereinheitlichen und das Datenschutzproblem lösen.

III.3. Zur Datenerhebung

Die wichtigsten Faktoren, die den Aufwand der Datenerhebung beeinflussen und bei einer Wiederholung des Verfahrens optimiert werden könnten, sind im Abschnitt B.I (S. 33 ff.) bereits geschildert worden. Im Folgenden werden einige weitere Aspekte der Datenerhebung diskutiert, die vorwiegend mit Hinblick auf die Aussagekraft der Daten von Bedeutung sind.

Die Erhebung wurde im Fach Chemie nach dem Prinzip „work done at“ durchgeführt. Das bedeutet, dass auch die Forschungsleistungen solcher Wissenschaftler, die innerhalb des Erhebungszeitraums eine Einrichtung verlassen, dieser Einrichtung für den Zeitraum vor dem Wechsel zugerechnet werden bzw. dort „verbleiben“ (vgl. A.III.3, S. 15 f.). Dies war der Tatsache angemessen, dass nicht die Leistung einzelner Wissenschaftler, sondern die Forschungsleistung einer ganzen Einrichtung bzw. Forschungseinheit erhoben und bewertet werden sollte; es war jedoch für manche Betroffene kontraintuitiv, da es der personenbezogenen Perspektive vieler gewohnter Begutachtungsprozesse zuwiderläuft. Bei der Alternative – der „current potential“-Erhebung[22], die sich nur auf die am Stichtag noch an der Einrichtung beschäftigten Wissenschaftler bezieht – ist jedoch die Gefahr von Zufallseffekten dadurch, dass Stellen am Stichtag unbesetzt sind, groß. Das britische Beispiel lehrt zudem, dass

[22] In der Pilotstudie Forschungsrating für das Fach Soziologie wurde die „current potential“-Erhebung gewählt.

„current potential“-basierte Bewertungen einen Anreiz für einen stichtagsbezogenen Transfermarkt erzeugen.

Bei beiden Erhebungsprinzipien ist die Fluktuation der Wissenschaftler relevant. Wenn eine Einrichtung innerhalb des Erhebungszeitraums von hoher Fluktuation betroffen ist oder aber im Erhebungszeitraum erst eingerichtet wurde, kann sich dies negativ auswirken. Daher wurden den Berichterstattern auch die Fluktuationsraten der Einrichtungen mitgeteilt. Eine hohe Fluktuation hat darüber hinaus den Nachteil, dass die Forschungsprodukte von Wissenschaftlern, die zum Stichtag nicht mehr an der Einrichtung tätig waren, häufig nicht mehr vollständig ermittelt werden konnten. Besonders im Falle emeritierter Wissenschaftler war dies problematisch. Nach Ansicht der Bewertungsgruppe Chemie ist es grundsätzlich kritisch, wenn eine Einrichtung das Wissen über das, was in den vergangenen Jahren mit ihren Mitteln geleistet wurde, mit dem Wechsel eines Wissenschaftlers verliert. Auch im Interesse der Rechenschaftslegung und als Entscheidungsgrundlage für die Einrichtungen selbst sollte solches Wissen unbedingt erhalten bleiben.

Der Bewertungsgruppe Chemie war bewusst, dass das „work done at“-Prinzip tendenziell weniger zukunftsgerichtete Aussagen erlaubt als das „current potential“-Prinzip. Es ist daher wichtig, durch eine gute Vorbereitung der Datenerhebung und -auswertung die Zeit zwischen dem Beginn des Erhebungszeitraums bzw. dem gewählten Stichtag und der Veröffentlichung der Ergebnisse bzw. der Bewertung der Daten zu verkürzen. Dazu werden folgende Empfehlungen abgegeben:

- Die Erhebungsdauer könnte dadurch verkürzt werden, dass den Einrichtungen eine längere *Vorlaufzeit* eingeräumt würde. Das Erhebungsprogramm, möglichst auch die Fragebogenformate, sollten den potentiell teilnehmenden Einrichtungen mit einem ausreichenden Vorlauf vor dem Stichtag bekannt gegeben werden, so dass die Daten im laufenden Betrieb gesammelt werden könnten und nicht im Nachhinein rekonstruiert werden müssten. Optimal wäre eine Vorlaufzeit entsprechend dem Erhebungszeitraum, sie sollte aber mindestens mehrere Monate betragen, besser 2-3 Jahre. Dies könnte nicht nur die Erhebungsdauer verkürzen, sondern auch den Erhebungsaufwand maßgeblich verringern.
- Der für die Datenerhebung gewählte *Erhebungszeitraum* betrug fünf Jahre. Diesen Zeitraum schätzt die Bewertungsgruppe Chemie zumindest für ihr Fach als angemessen ein. Eine Verkürzung des Erhebungszeitraums würde die Datenmenge sowie die Notwendigkeit, Daten über mehrere Jahre vorzuhalten, und damit den

Erhebungsaufwand verringern. Auch die „Gefahr“ hoher Fluktuationsraten wäre damit geringer. Ein kürzerer Erhebungszeitraum würde jedoch die Gefahr mit sich bringen, dass Schwankungen in der Forschungsleistung nicht ausgeglichen werden und die Resultate mehrjähriger Projekte nicht angemessen gewürdigt werden können. Zudem ist für die Zitationsanalysen ein ausreichender Zitationszeitraum notwendig, der in den Naturwissenschaften in der Regel mindestens drei Jahre betragen sollte.[23] Bei einer Verkürzung des Erhebungszeitraums könnte dies u. U. nur für die Publikationen eines einzigen Jahrgangs erreicht werden.

Deutlichen Optimierungsbedarf gibt es bei der Gestaltung und technischen Realisierung der Fragebögen. Häufig haben Fachkoordinatoren Schwierigkeiten bei der Übertragung bereits vorhandener Daten in die teilweise geschützten Formulare, die Trennung in tabellarische und textförmige Antwortmöglichkeiten sowie inhaltlich begründete Beschränkungen etwa von Listen auf eine bestimmte Anzahl von Einträgen bemängelt. Gerade diese technischen Umstände haben teilweise zu Verstimmungen und einer negativen Grundhaltung gegenüber der durch das Rating verursachten zusätzlichen Arbeitsbelastung beigetragen. Bei einer Wiederholung oder Verstetigung des Verfahrens sollte die technische Realisierung des Erhebungsinstruments und die Datensammlung einem spezialisierten Anbieter übertragen werden. Die inhaltliche Vorbereitung der Erhebung sowie die Kontrolle und Aufbereitung der Daten sollte indes weiterhin einer Geschäftsstelle obliegen, die das Verfahren organisiert und mit den Gutachtern und den Einrichtungen kooperiert. Eine denkbare Vereinfachung der Erhebung wäre ein Online-Erhebungsinstrument. Dies war im Rahmen der Pilotstudie aus zeitlichen Gründen nicht möglich, wäre aber ggf. für eine Verstetigung des Verfahrens erneut auf seine Machbarkeit zu überprüfen. Zu den zu lösenden Problemen gehört unter anderem die Verbindung von Text- und Tabellenteil der Fragebögen unter Wahrung der „copy and paste“-Möglichkeit, die Erhebung auf zwei Aggregationsebenen und eine hinreichend flexible Anwendung mit mehreren Nutzern innerhalb der bewerteten Einrichtungen.

Die Erfahrungen aus der Pilotstudie haben außerdem gezeigt, dass die möglichst präzise, aber zugleich nicht zu einengende Formulierung der Fragen für die Homogenität und Vergleichbarkeit der Daten einerseits und die individuellen Ausdrucksmöglichkeiten der Einrichtungen andererseits entscheidend ist. Zu offene Fragen führen zu sehr heterogenen Angaben, zu eng definierte Fragen provozieren Kritik,

[23] Vgl. van Raan: „Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises“, in: Scientometrics 36,3 (1996), 397-420, S. 403.

wenn Einrichtungen glauben, sich nicht adäquat darstellen zu können. Hier wird es jeweils Aufgabe der Bewertungsgruppen sein, in der fachspezifischen Entwicklung der Fragebögen das richtige Maß zu finden und damit zur Sicherung einer hohen Qualität der Datenbasis beizutragen (zu einzelnen Indikatoren s. C.I.).

III.4. Zur Publikationserhebung und Zitationsanalyse

Die bibliometrische Analyse basierte auf der Datenbank Web of Science (WoS) des Anbieters Thomson Scientific. Zum gegenwärtigen Zeitpunkt ist dies nach Einschätzung der meisten Experten die verlässlichste Datenbasis für bibliometrische Analysen.[24] Bei einer etwaigen Wiederholung des Verfahrens im Fach Chemie wird zu prüfen sein, ob die derzeit in der Entwicklung befindlichen Konkurrenzprodukte bis dahin an Zuverlässigkeit gleichziehen oder sogar Vorzüge hinsichtlich Abdeckung und Indikatorenbestand geltend machen können. Ein stichprobenartiger Vergleich der Daten verschiedener Anbieter wäre als Entscheidungsgrundlage zu empfehlen.

Die Einrichtungen wurden gebeten, die in der Datenbank auf Basis der Namen der von ihnen benannten Wissenschaftler und der Adressen der Einrichtungen recherchierten Publikationen zu überprüfen und ggf. zu ergänzen. Das IWT hat durch die Suche nach zahlreichen alternativen Schreibweisen der Einrichtungen sowie der einzelnen Autoren diese bekannte Fehlerquelle nach Möglichkeit verringert. Dennoch war die Rückkoppelungsschleife für eine möglichst vollständige und korrekte Erfassung der Publikationen unverzichtbar. Nur gemeinsam mit den Einrichtungen kann eine ausreichend zuverlässige Publikationserfassung sichergestellt werden, um auf dieser Basis dann eine bibliometrische Analyse vorzunehmen.

Bei den verwendeten Indikatoren handelt es sich um etablierte Messgrößen, die in der bibliometrischen Forschung anerkannt und in ihrem Verhalten gut untersucht sind. Die Möglichkeit, die Berechnung der einzelnen Werte anhand der Rohdaten im Einzelfall nachvollziehen zu können, spielte für das Vertrauen der Gutachter in die Daten wie für die Korrektur von Verzerrungen oder Fehlern eine entscheidende Rolle. Bei einem potentiell folgenreichen Verfahren wie dem Forschungsrating sind deshalb innovative Indikatoren nur mit Vorbehalt zu benutzen. So wurde für die Pilotstudie auch der sogenannte Hirsch-Index der einzelnen Forschungseinheiten[25] berechnet. Der „h-Index“ wurde ursprünglich für die Bewertung der in Zitations- und Publika-

---

[24] Vgl. Higher Education Funding Council for England (HEFCE): Research Excellence Framework: Consultation on the assessment and funding of higher education research post-2008, November 2007/34, S. 10.

[25] Hirsch. J.E. (2005): An index to quantify an individual’s scientific research output. arXiv:physics/0508025v5, 29 Sep 2005.

tionsleistung bemessenen wissenschaftlichen Lebensleistung von Einzelpersonen konzipiert und ist insofern nicht einfach auf Forschungseinheiten übertragbar. Vor einer Anwendung dieses Indikators im Rahmen eines auf Forschungseinheiten bezogenen Forschungsratings müsste zunächst geklärt werden, wie sich a) die unterschiedliche Größe der Einheiten, b) die zeitliche Einschränkung auf einen bestimmten Bewertungszeitraum und c) die unterschiedliche Spezialisierung der Forschungseinheiten auf diesen Indikator auswirkt. Da diese Fragen zum Zeitpunkt der Bewertung noch nicht geklärt waren, wurde der „h-Index" der Forschungseinheiten den Gutachtern nicht mitgeteilt und daher auch nicht verwendet. Trotzdem weist der h-Index von allen berechneten Indikatoren gemeinsam mit dem Indikator ZP/FCS_m die höchste Korrelation mit der Bewertung der Forschungsqualität (Kriterium I) auf. In Zukunft sollte die Weiterentwicklung dieses und anderer innovativer Indikatoren deshalb aufmerksam verfolgt werden.

Die verwendeten Indikatoren wurden von der Bewertungsgruppe als sehr belastbar angesehen. In der Bewertungspraxis zeigte sich, dass kein einzelner Indikator isoliert betrachtet ein zuverlässiges Qualitätsurteil erlaubt. Erst in der Gesamtschau der Indikatoren und unter Berücksichtigung der von den Einrichtungen bereitgestellten Kontextinformationen konnte eine faire und verlässliche Bewertung vorgenommen werden. Durch dieses Vorgehen konnte außerdem der bei einer rein bibliometrischen Auswertung bestehenden Sorge entgegengewirkt werden, risikoreiche und innovative Forschung würde benachteiligt. Die gleichzeitige Verwendung mehrerer Indikatoren einschließlich solcher, die auf einen internationalen Mittelwert normiert sind, bietet zudem am ehesten die Chance, Manipulationsstrategien zu erkennen.[26]

Die bibliometrische Analyse war für das Forschungsrating Chemie wesentlich. Bei einer Wiederholung des Forschungsratings ist eine bibliometrische Analyse, ggf. unter Fortentwicklung der Methode, unverzichtbar. Die Bewertungsgruppe spricht sich jedoch aufgrund der gemachten Erfahrungen und der Komplexität der bibliometrischen Indikatoren nachdrücklich gegen eine rein indikatorenbasierte Bewertung aus.

[26] Das Centre for Science and Technology Studies der Universität Leiden (CWTS) hat eine ausführliche Analyse der Folgen von Leistungsmessung anhand bibliometrischer Indikatoren beim RAE in den Jahren 1992 bis 2001 durchgeführt und „Anpassungsstrategien" im Publikations- und Zitierverhalten der betroffenen Wissenschaftler festgestellt. Als Reaktion darauf wird empfohlen, keine absoluten Publikationswerte zu verwenden und Impactfaktoren von Zeitschriften nicht zu berücksichtigen. Der Indikator Zitationen pro Publikation bezogen auf den Fachgebietsdurchschnitt wird als der am wenigsten manipulationsanfällige Indikator beschrieben. Vgl. CWTS: Scoping study on the use of bibliometric analysis to measure the quality of research in UK higher education institutions, Report to HEFCE by the Centre for Science and Technology Studies, Leiden University, November 2007, S. 34ff.

III.5. Zur Datenaufbereitung

Die Geschäftsstelle hat alle erhobenen Daten darauf überprüft, ob die Erhebungsregeln eingehalten wurden und die Angaben in sich konsistent waren. Die häufigsten Gründe für Rückfragen waren lückenhafte Angaben sowie offenkundige Verletzungen teils trivialer Regeln, die in den Fragebögen explizit formuliert waren – der häufigste Fehler war, dass Angaben zu Aktivitäten gemacht wurden, die nicht in den Erhebungszeitraum fielen (vgl. A.III.3, S. 15 f.). Auch wenn diese Probleme mit einer Routinisierung und steigenden Akzeptanz eines solchen Verfahrens geringer werden, zeigt dies, wie wichtig auf der einen Seite ein komfortables und übersichtliches Erhebungsinstrument, auf der anderen Seite aber auch eine sorgfältige Datenkontrolle ist.

Bei einzelnen Fragen (bspw. bei der Angabe von Plenarvorträgen auf internationalen Kongressen und „named lectures“) gab es fachintern offenkundig Auffassungsunterschiede, die seitens der Ausfüllenden zu einem bisweilen ausufernden Meldeverhalten führten. Soweit möglich, wurde versucht, dies durch Präzisierungen der Regeln im laufenden Verfahren (bspw. durch die Beschränkung auf exemplarische Angaben) zu korrigieren. Die abschließende Datenkontrolle konnte solche inhaltlichen Missverständnisse nur begrenzt bereinigen, da es dazu einer innerfachlichen Entscheidung bedurft hätte.

Nach der Kontrolle der Daten und den fälligen Nacherhebungen wurden die Ergebnisse in Datenberichten für jede Einrichtung zusammengefasst, in denen die Rahmeninformationen und die Indikatoren in einen Sinnzusammenhang gemäß der definierten sechs Kriterien gestellt wurden. Diese Datenberichte wurden den Einrichtungen zur Abschlusskontrolle vorgelegt. Allerdings hatte eine große Kulanz in der Erhebungsfrist dazu geführt, dass diese Rückkopplungsschleife sich in einigen Fällen auf eine Woche verkürzte. Dieser Rückkopplung muss künftig mehr Zeit eingeräumt werden.

Erst im Anschluss an die Rückkopplung und nach Einarbeitung evtl. notwendiger Änderungen konnten die quantitativen Daten für die Grundgesamtheit ausgewertet werden und Lagemaße (genauer: Perzentilwerte) ebenfalls in die Datenberichte eingefügt werden. Weitere Verteilungsmaße (1. und 3. Quartil, Median) wurden in dem „Leitfaden zu den Datenberichten“[27] mitgeteilt und durch statistische Analysen er-

[27] Dieser ist einzusehen unter www.wissenschaftsrat.de/pilot_start.htm.

gänzt. Nach Ansicht der Bewertungsgruppe Chemie hat sich die Umrechnung der quantitativen Daten in skaleninvariante Lagemaße grundsätzlich bewährt. Allerdings ist darauf hinzuweisen, dass solche Werte im Bewertungsprozess eine Komplexitätsreduktion darstellen, die unbedingt reflektiert werden sollte. Beispielsweise kann es vorkommen, dass die Signifikanz der Perzentildifferenzen überschätzt wird, wenn die im Leitfaden enthaltenen statistischen Analysen, die auch die Streuung der zugrunde liegenden Messwerte wiedergeben, nicht mit berücksichtigt werden. Auch aus diesem Grunde warnt die Bewertungsgruppe Chemie vor der Berechnung hoch aggregierter Zahlen, etwa gewichteter Gesamtindikatoren für einzelne Kriterien, da sie das vorhandene und für die Bewertung relevante Differenzierungspotential der einzelnen Daten zu weit reduzieren. Es wäre zu prüfen, ob es ein für einen Bewertungsvorgang noch besser geeignetes Maß für die Position eines Wertes innerhalb einer bestimmten Grundgesamtheit gibt als den Perzentilwert.

Insgesamt wurde die Aufbereitung der Daten in Datenberichten sowohl von der Bewertungsgruppe als auch von einigen teilnehmenden Einrichtungen, die dazu Stellung genommen haben, positiv beurteilt.

III.6. Zur Bewertungsphase

Die Bewertung wurde durch eine 15-köpfige, international besetzte Gruppe von Fachvertretern vorgenommen, die die meisten Teilgebiete der Chemie mit ihrer Expertise abdeckte. Allerdings war in Einzelfällen, etwa kleinen Spezialgebieten der Chemie, oder bei Befangenheiten der Rückgriff auf externe Sondergutachter notwendig. Dies war besonders dann hilfreich, wenn ein Sondergutachter jeweils mehrere Forschungseinheiten zu bewerten hatte und somit einen internen Vergleichsmaßstab erhielt. Einzelgutachten von ansonsten nicht involvierten Wissenschaftlern waren mangels dieses Vergleichsmaßstabs schwieriger in den Prozess zu integrieren. Es ist daher empfehlenswert, auf wenige Sondergutachter zurückzugreifen, die dann jeweils für mehrere Bewertungen zuständig sein sollten.

Die Bewertungsphase war in eine individuelle und eine plenare Bewertungsphase aufgeteilt. Für die Ergebnisse der individuellen Bewertungen wurden keine Vorgaben – wie etwa eine bestimmte Verteilung auf die Skalenstufen – gemacht. Auch die Gewichtung der einzelnen Bewertungsaspekte bei der Bildung einer Gesamtnote zu einem bestimmten Kriterium blieb dem einzelnen Berichterstatter überlassen. Der recht hohe Konsens der Berichterstatter (s. o. Tabelle 3: Übereinstimmung der

Gutachterurteile) lässt darauf schließen, dass die vorangegangenen Diskussionen über die Operationalisierung der Kriterien dazu geführt hatten, ein gemeinsames Verständnis der Interpretation der einzelnen Indikatoren auszubilden. Es erscheint daher nicht notwendig, den Bewertungsprozess stärker zu reglementieren.

Die fünfstufige Bewertungsskala entspricht internationalen Beispielen und hat sich im Wesentlichen bewährt. Die Notenstufen mit Ausnahme der Spitzennote weder semantisch noch durch eine vorgegebene Verteilung näher zu bestimmen, war im Bewertungsprozess unproblematisch und würde es bei einer Ausweitung des Verfahrens auf weitere Fächer erleichtern, einen einheitlichen Rahmen vorzugeben. Dass die Skala im Kriterium Wissensvermittlung und -verbreitung auf drei Stufen reduziert wurde, schränkt die Vergleichbarkeit der Kriterien ein, war aber des heterogenen und lückenhaften Datenmaterials wegen unumgänglich. Nach Ansicht der Bewertungsgruppe Chemie wäre es wünschenswert, künftig auf solche Abweichungen verzichten zu können. Ziel sollte eine einheitliche Bewertungsskala sein. Denkbar wäre es allerdings, für bestimmte Kriterien eine asymmetrische Skala zu verwenden, wie dies für das Kriterium Forschungsqualität durch Einfügen der Zwischennote „sehr gut bis exzellent“ der Fall war, wenn die Differenzierung des Datenmaterials im unteren Bereich wenig ausgeprägt und daher eine Differenzierung der Bewertung entsprechend schwierig ist.

Die Grundlage der Bewertung waren die Datenberichte. Jeder Gutacher erhielt auch die Datenberichte derjenigen Einrichtungen, für die er nicht als Berichterstatter zuständig war. Dies stellte sicher, dass in der plenaren Bewertung alle Mitglieder der Bewertungsgruppe auf der gleichen Informationsbasis diskutieren konnten. Die Datenberichte wurden in der vorgelegten Form von den Gutachtern als übersichtlich und gut strukturiert empfunden. In den Bewertungssitzungen wurden nach Bedarf zusätzliche Auswertungen vorgelegt. In der abschließenden Sitzung, die der Konsolidierung der Bewertungen und der Abstimmung über die „Extremnoten“ diente, wurden vergleichende Übersichten erstellt, die nach den bis dahin erzielten Bewertungsergebnissen sortiert waren. Dies erwies sich als hilfreiches Mittel, diskussionswürdige Fälle aufzufinden.

Kommentare zu den Bewertungen für die Öffentlichkeit oder für die betroffene Einrichtung wurden nur in Einzelfällen festgehalten, etwa dann, wenn die Bewertung durch außergewöhnliche Bedingungen stark beeinflusst wurde, oder um zu signalisieren, dass die beobachtbaren Entwicklungen eine positive Tendenz für die Zukunft

erwarten ließen. Auch die Klassifizierung als „nicht bewertbar“ wurde individuell kommentiert. Die Option, zu jeder Einrichtung einen Kommentar zu verfassen, wurde verworfen, da das Instrument der Kommentierung ausschließlich als Hinweis auf Besonderheiten benutzt werden sollte. Weitergehende Analysen, insbesondere Vermutungen darüber, was die Leistungsfähigkeit der bewerteten Einrichtung erklärt, sind ohne eine umfassendere Einzelevaluation nicht zu substantiieren.

In der Bewertung zeigte sich eine unterschiedlich hohe Belastbarkeit der einzelnen Kriterien. Die Bewertung der Kriterien Forschungsqualität und Impact/Effektivität[28] war belastbarer als die Bewertung der übrigen Kriterien. Am wenigsten belastbar und differenziert war die Bewertung des Kriteriums Wissensvermittlung und -verbreitung, vor allem aufgrund einer schmalen und heterogenen Datengrundlage, die nur einen (experimentellen) quantitativen Indikator umfasste (vgl. Anhang, C.VI, S. 64).

Insgesamt wird das Bewertungsverfahren durch die Bewertungsgruppe Chemie positiv beurteilt. Sie spricht sich daher dafür aus, das gewählte Verfahren in seinen Grundzügen beizubehalten, wenn das Forschungsrating wiederholt werden sollte.

III.7. Zur Bewertungsmatrix

Die Bewertungsmatrix fasste das Ergebnis der Operationalisierung zusammen und setzte den Rahmen sowohl für die Datenerhebung als auch für die Bewertungsphase. Der fachspezifischen Ausarbeitung einer solchen Matrix ist daher besonderes Gewicht beizumessen.

Die Bewertungsmatrix kann ex post kontrolliert werden, indem der Zusammenhang zwischen dem Bewertungsergebnis und den zugrunde liegenden quantitativen Indikatoren analysiert wird. Eine solche Auswertung sollte auch bei einer Verstetigung des Verfahrens regelmäßig vorgenommen werden, um ggf. die Indikatorensets für einzelne Fächer fortzuschreiben. Dabei sollte indes berücksichtigt werden, dass auch bei hoher Korrelation zwischen Bewertung und zugrunde gelegtem quantitativem Indikator Ausnahmen möglich sind, bei denen die Gutachter in der Bewertung aufgrund von qualitativen Informationen zum Einzelfall bewusst von den quantitativen Indikatoren abweichen. In solchen Fällen kann ein der Besonderheit des Einzelfalls angemessener, bezogen auf die Gesamtheit der bewerteten Einrichtungen nur schwach mit dem Bewertungsergebnis korrelierter Indikator für das Gutachterurteil ausschlaggebend sein. Die Relevanz einzelner Indikatoren kann also nur bestimmt

[28] Unter „Impact/Effektivität“ ist die Sichtbarkeit der Forschung in der scientific community zu verstehen.

werden, wenn neben den Ergebnissen einer Korrelationsanalyse auch die Aussagen der Gutachter zu einzelnen Indikatoren zugrunde gelegt werden.

Für die Pilotstudie spiegeln die Befunde der Korrelationsanalyse für die quantitativen Indikatoren die in den plenaren Bewertungssitzungen geführten Diskussionen über die Aussagekraft und Belastbarkeit der Indikatoren im Wesentlichen wider.

Einige quantitative Indikatoren korrelieren wenig oder überhaupt nicht mit der Bewertung des entsprechenden Kriteriums. Daraus könnten Vereinfachungen für die Datenerhebung abgeleitet werden, indem diese nur gering oder überhaupt nicht korrelierten Daten kritisch hinterfragt und ggf. künftig nicht mehr erhoben bzw. berechnet werden. Bevor eine solche Schlussfolgerung gezogen wird, ist allerdings zu prüfen, ob die geringe Korrelation wirklich mit der inhaltlichen Irrelevanz des jeweiligen Indikators zu erklären ist oder doch eher mit einer geringen Reliabilität der bislang verfügbaren Daten. In dem Fall wäre ein Verzicht auf den Indikator möglicherweise der falsche Ansatz, es sollte stattdessen erwogen werden, wie dieses Datum künftig verlässlicher oder einfacher zu erheben wäre.[29]

Insgesamt ist nach Ansicht der Bewertungsgruppe Chemie die Indikatorenbasis für die Dimension Wissenstransfer der kritischste Punkt der Bewertungsmatrix. Teilweise konnten die notwendigen Daten nicht erhoben werden, teilweise waren die Angaben der Einrichtungen stark heterogen und daher schwer einheitlich zu bewerten. Zur Verbesserung dieser Situation sind zwei Optionen denkbar:

1. Zusammenlegung der Kriterien V Transfer in andere gesellschaftliche Bereiche und VI Wissensvermittlung und -verbreitung der Dimension Wissenstransfer zu einem Kriterium. Dadurch Reduktion Anzahl der Kriterien (6 auf 5).
2. Beibehaltung beider Kriterien bei schärferer Differenzierung ihrer Bedeutung verbunden mit Neuentwicklung von Indikatoren für die Dimension Wissenstransfer.

Eine Zusammenlegung der beiden Kriterien der Dimension Wissenstransfer zu einem Kriterium hätte Nachteile für solche Einrichtungen, deren Stärke im anwendungsorientierten Bereich liegt. Stärke im anwendungsorientierten Bereich ist nicht mit Stärke in der Wissensvermittlung gleichzusetzen, was auch die geringe Korrelation der beiden Kriterien bestätigt. Ein Verzicht auf diese Differenzierung des Profils hinsichtlich des Wissenstransfers wäre insofern nicht optimal. Neue Indikatoren für die Dimension Wissenstransfer müssten in enger Kooperation und Abstimmung mit

[29] Zur Diskussion der einzelnen Kriterien und der zugeordneten Indikatoren vgl. Teil C.

den Fachgesellschaften verschiedener naturwissenschaftlicher Fächer entwickelt werden. Die Bewertungsgruppe empfiehlt, die beiden Kriterien weiterhin getrennt zu bewerten, dabei aber deutlicher zu differenzieren. Kriterium V sollte in der Chemie ausschließlich den Transfer in die Wirtschaft betreffen, unter Kriterium VI würde dann die Vermittlung von Wissen in die breitere Öffentlichkeit bewertet. Dies würde eine differenziertere Bewertung erlauben und bei den Einrichtungen das Verständnis dafür erhöhen, welche Inhalte für welches Kriterium bewertungsrelevant sind. Zugleich sollten die Indikatoren für Kriterium VI modifiziert werden (s.u. C.VI.).

Trotz relativ großer Zusammenhänge zwischen der quantitativen Datenbasis und der Bewertung ist festzuhalten, dass kein Kriterium ausschließlich über die zugrundeliegenden quantitativen Daten bewertet wurde, sondern immer die Zusammenstellung aller relevanten (auch qualitativen) Indikatoren die Bewertung bestimmte. Auch unter den qualitativen Indikatoren wurden einige im Bewertungsvorgang für weniger relevant oder weniger reliabel erachtet. Details und Empfehlungen dazu finden sich in Teil C dieses Berichts.

Wie in Teil A.IV.2 näher erläutert, korreliert keines der Kriterien so hoch mit einem anderen Kriterium, dass es redundant wäre. Insgesamt führen die sechs Kriterien zu einer sinnvollen Differenzierung der Bewertung, so dass es auch keine dominanten „Cluster“ in der Bewertung der Einrichtungen gibt, sondern die Bewertungsergebnisse insgesamt stark heterogen sind und damit offenbar das jeweils individuelle Leistungsprofil der einzelnen Einrichtungen adäquat erfassen. Die in diesem Verfahren erprobte mehrdimensionale Bewertung hat sich somit bewährt und sollte auch für eine Weiterentwicklung beibehalten werden.

III.8. Zur Veröffentlichung

Die gewählte Veröffentlichungsform in einer Pressekonferenz verbunden mit einem Pressegespräch, an dem auch Vertreter von GDCh und VCI teilnahmen, sowie über einen begleitenden Internet-Auftritt ermöglichte es, eine breite Öffentlichkeit zu erreichen. Für die Akzeptanz des Verfahrens in der Wissenschaft war es hilfreich, dass die Einrichtungen ihre eigenen Ergebnisse und den endgültigen Datenbericht mit statistischen Informationen zu den sie betreffenden Daten etwa eine Woche vor der Veröffentlichung erhalten haben.

Reaktionen auf die Veröffentlichung des Forschungsratings kamen von verschiedenen Seiten: Erstens von der Presse, wo unterschieden werden kann zwischen Mel-

dungen, die hauptsächlich die Ergebnisse wiedergeben und solchen Meldungen, die auch das Verfahren erläutern oder sich sogar kritisch – positiv – damit auseinandersetzen. Zweitens wurden Meldungen von den teilnehmenden Einrichtungen selbst veröffentlicht und drittens von den Trägern (Länder oder Wissenschaftsorganisationen) der teilnehmenden Einrichtungen. Insgesamt sind die Reaktionen überwiegend positiv. Wichtige überregionale Printmedien (ZEIT, FAZ, FR) begrüßen das Verfahren. Dass die Universitäten, die gut abgeschnitten haben, sich nicht negativ über das Verfahren äußern würden, war erwartbar. Die einzige negative Reaktion war die Pressemitteilung der MPG, die sich kritisch zur Aufwand-Nutzen-Relation äußert. Dass die Pressearbeit insgesamt gelungen ist, lässt sich auch daran erkennen, dass im Presseecho keine gravierenden Missverständnisse bzgl. des Verfahrens zu finden sind.

Steuerungsgruppe und Bewertungsgruppe haben Wert darauf gelegt, dass nicht nur die Ergebnisse selbst veröffentlicht wurden, sondern bereits eine erste Auseinandersetzung mit dem Verfahren stattfand. Der veröffentlichte Bericht der Bewertungsgruppe Chemie (s. Anlage) enthielt neben Aussagen zu Stärken und Schwächen der Chemie in Deutschland deshalb auch eine Schilderung des Verfahrens, in der auf die wesentlichen methodischen Diskussionspunkte eingegangen wurde. Empfehlungen zum Verfahren des Forschungsratings wurden hingegen noch nicht ausgesprochen, da die Adressaten dieser Veröffentlichung in erster Linie die „scientific community“ des Faches Chemie sowie Entscheidungsträger in Universitäten, außeruniversitären Einrichtungen, den Ländern und bei den Trägerorganisationen waren und nur in zweiter Linie Experten für Verfahren der Forschungsbewertung.

Die Visualisierung der Ergebnisse der einzelnen Einrichtungen in Gestalt von Balkendiagrammen war grundsätzlich geeignet, die Bewertungen in ihrer Differenziertheit angemessen widerzuspiegeln. Die Darstellungen waren jedoch für Leser, die mit dem Verfahren nicht vertraut waren, bereits relativ komplex, u. a. durch die Verwendung zweier unterschiedlicher Bewertungsskalen für die Kriterien I bis V respektive VI. Für Zwecke der Pressearbeit erstellte die Geschäftsstelle deshalb Übersichtstabellen, die wahlweise eine Sortierung der Einrichtungen nach einem der sechs Kriterien erlaubten. Diese tabellarischen Übersichten wurden von der Presse gern angenommen. Durch das Bereitstellen von einfachen Übersichten ist es in gewissem Umfang möglich, die Art und Weise der Komplexitätsreduktion in der Presse zu steuern und dadurch Fehlern und größerer Willkür vorzubeugen.

Bei einer Verstetigung eines Forschungsratings sollte großer Wert auf eine ansprechende und verständliche graphische Darstellung der Ergebnisse gelegt werden. Hierfür ist ggf. die Unterstützung eines professionellen Grafikbüros zu suchen.

Es sollte künftig unbedingt möglich sein, die auf der niedrigen Ebene der Forschungseinheiten vorgenommenen Bewertungen der Forschungsqualität auch der Öffentlichkeit zugänglich zu machen. Im Rahmen der Pilotstudie wurden diese Details nur den Einrichtungen intern zur Verfügung gestellt, da die Veröffentlichung personenbeziehbarer Bewertungen ohne Zustimmung der Betroffenen datenschutzrechtlich bedenklich ist. Weil diese Details indes einen hohen Informationsgehalt haben, der möglicherweise auch für externe Adressaten von Interesse sein könnte, etwa als Orientierungshilfe für zukünftige Doktoranden oder für die Industrie, die Ansprechpartner zu einzelnen Problemfeldern sucht, sollten die Forschungseinheiten künftig in der Regel so zugeschnitten sein, dass eine Personenbeziehbarkeit der Bewertung ausgeschlossen werden kann (vgl. III.2, S. 41 f.). Wo dies ausnahmsweise nicht gelingt, sollten die Einrichtungen von vorneherein darauf aufmerksam gemacht werden, dass die Bewertungen der Forschungseinheiten veröffentlicht werden. Unter diesen Voraussetzungen kann den Einrichtungen zugemutet werden, die betroffenen Wissenschaftler um ihr Einverständnis zu bitten.

B.IV. Zum weiteren Vorgehen

Die Bewertungsgruppe Chemie ist überzeugt, dass das Forschungsrating geeignet ist, seinen verschiedenen Adressaten eine wertvolle Grundlage für verschiedene Entscheidungen zu liefern. Die Transparenz, die das Verfahren schafft, kann den Qualitätswettbewerb fördern und den Einrichtungen helfen, ihre Profile auf einer verlässlichen Basis weiterzuentwickeln und damit ihre Qualität zu steigern. Das Fach Chemie insgesamt profitiert von einer größeren Sichtbarkeit auch im internationalen Raum. Das aufwendige „Informed Peer Review“ ist einer rein quantitativen Datenauswertung deutlich überlegen, die Mehrdimensionalität der Bewertung ist der Komplexität moderner Wissenschaft angemessen und die Berücksichtigung sowohl der universitären als auch der außeruniversitären Forschung liefert (in der Pilotstudie erstmals) einen adäquaten Überblick über die differenzierte Forschungslandschaft in Deutschland.

Für die bewerteten Einrichtungen, insbesondere für die Universitäten, wären die Ergebnisse der Pilotstudie noch hilfreicher, wenn auch für die benachbarten Fächer

vergleichbare Bewertungen vorlägen. Die Bewertungsgruppe Chemie spricht sich deshalb dafür aus, das Verfahren möglichst bald auszuweiten und zumindest die naturwissenschaftlichen Nachbardisziplinen der Chemie, Physik und Biologie, ebenfalls in Form eines Ratings zu bewerten. Zu diesem Zweck sollte zunächst eine umfassende Taxonomie der in Frage kommenden Fächer entwickelt werden. Dadurch würden auch Zuordnungsschwierigkeiten an den Rändern der Disziplinen und bei interdisziplinärer Forschung verringert werden. Dabei müsste im Verfahren sichergestellt sein, dass der Zuschnitt der jeweils gemeldeten Forschungseinheiten der einzelnen Einrichtungen stärker standardisiert wäre.

Der Nutzen des Forschungsratings würde noch einmal erheblich steigen, wenn ein Fach nach einigen Jahren noch einmal bewertet wird, so dass die bewerteten Einrichtungen aus den Ergebnissen Entwicklungstendenzen ablesen können. Auf Basis solcher Informationen könnten dann auch weitreichende strategische Entscheidungen besser begründet werden. Die Bewertungsgruppe Chemie spricht sich unter dem Vorbehalt der Berücksichtigung der in B.I. und B.II. genannten Verbesserungsvorschläge deshalb dafür aus, in etwa fünf Jahren erneut ein Forschungsrating der Chemie vorzunehmen. Ein Mitglied der Bewertungsgruppe lehnt eine Wiederholung oder Ausweitung des Verfahrens angesichts des Aufwands, ein weiteres Mitglied grundsätzlich ab, da seiner Ansicht nach objektive und zuverlässige Kriterien für die Bewertung der Qualität von Forschung fehlen und derartige Evaluationen einen innovationsfeindlichen Zwang zu „mainstream“-Forschung ausüben.

Eine direkte Verknüpfung mit der staatlichen Mittelvergabe sollte es auch bei einer Wiederholung des Verfahrens nicht geben, da dies der Autonomie der Einrichtungen zuwiderlaufen würde und die Gefahr mit sich brächte, dass der Wettbewerb zwischen den Einrichtungen sich auf die Erfüllung von Ratingkriterien verengt.

C. Anhang: Empfehlungen zu den einzelnen Kriterien und zur Datengrundlage der Pilotstudie Forschungsrating Chemie

Im Rahmen der Pilotstudie ist bewusst eine relativ umfangreiche Datenbasis erhoben worden, um sicherzustellen, dass auf jeden Fall ausreichend Informationen für eine belastbare und differenzierte Bewertung vorliegen. Es gehört deshalb zu den Lehren, die aus der Pilotstudie zu ziehen sind, welche Indikatoren sich letztlich nicht in hinreichend belastbarer Qualität erheben ließen oder aufgrund von Interpretationsschwierigkeiten kaum berücksichtigt wurden. Solche Daten sollten künftig nicht mehr erhoben werden. Ein Indiz für die Bedeutung der einzelnen quantitativen Indikatoren kann die Korrelation mit den Bewertungsergebnissen sein. Da bestimmte Indikatoren aufgrund von qualitativen Kontextfaktoren nur in Einzelfällen hinzugezogen wurden, kann aber letztlich nur von den Gutachtern beurteilt werden, welche Indikatoren verzichtbar sind.

C.I. Kriterium I, Forschungsqualität

Das Kriterium Forschungsqualität wurde als das für die Bewertung der Forschungsleistung zentrale Kriterium angesehen. Daher wurde beschlossen, hier eine Differenzierung innerhalb einer Einrichtung durch die Einführung der Aggregationsebene Forschungseinheit zu ermöglichen. Zwar war der Aufwand für die Erhebung auf dieser Aggregationsebene teils recht hoch, die erhobenen Daten lieferten aber eine sehr belastbare Bewertungsgrundlage. Das Ziel, eine differenzierte und aussagekräftige Bewertung der Forschungsqualität vorzunehmen, konnte erreicht werden.

Alle in der Bewertungsmatrix enthaltenen quantitativen Indikatoren sind statistisch signifikant mit der Bewertung der Forschungsqualität korreliert. Das Vertrauen der Gutachter in die quantitative Datenbasis war bei diesem Kriterium sehr groß, alle erhobenen quantitativen Indikatoren wurden in der Bewertung berücksichtigt. Dies machte sich nicht nur in den plenaren Diskussionen bemerkbar, sondern schlägt sich auch in ihrer signifikant (positiven) Korrelation mit dem Bewertungsergebnis nieder. Die verwendeten quantitativen wie qualitativen Indikatoren sollten bei künftigen Bewertungen wieder verwendet werden, wobei folgende Empfehlungen zu berücksichtigen sind:

- *Publikationslisten*: Für die Urteilsbildung der Gutachter ist der Zugriff auf die standardisierten Publikationslisten, die gemeinsam mit den bewerteten Einrichtungen bereinigt wurden, auch künftig unverzichtbar.

- *normierte Zitationszahlen (ZP/FCS$_m$)*: Der auf das Fachgebiet normierte Indikator ZP/FCS$_m$ war für die Gutachter in einzelnen Fällen schwierig zu interpretieren, da die Publikationen den im WoS definierten Teilgebieten zugeordnet werden mussten, wobei die Zuordnung nicht für jede Publikation individuell vorgenommen wurde, sondern sich nach der Zeitschrift richtete, in der sie erschienen war. Dennoch ist diese Normierung sehr hilfreich. In einigen Fällen, etwa bei Teilgebieten, die in der WoS-Klassifikation nicht adäquat abgebildet sind, oder bei sehr heterogenen Forschungseinheiten haben die Gutachter zusätzlich die Rohdaten der Zitationsanalyse zur Bewertung herangezogen, die die Fachgebietsnormierung für jede Publikation einzeln auswiesen. Diese Informationen waren ein wichtiges Korrektiv und sollten auch künftig im Bedarfsfall für die Gutachter verfügbar gemacht werden.
- *Zitationszahlen*: Die Bewertungsgruppe hat zur Einschätzung des Zitationserfolgs auch die absoluten Zitationszahlen verwendet. Sie empfiehlt deshalb, anders als in der Pilotstudie, künftig die absolute Zahl der Zitationen, in den Datenberichten direkt auszuweisen.
- *Drittmittelprojekte*: Die Erhebung von Drittmittelprojekten auf Ebene der Forschungseinheiten, insbesondere die Erfassung der Bewilligungssummen einzelner Drittmittelprojekte, war vor allem wegen des langen Erhebungszeitraums und der retrospektiven Erhebung aufwendig. Die Erhebungsprobleme waren allerdings von Einrichtung zu Einrichtung sehr unterschiedlich und schienen vor allem von der Drittmittelverwaltung und dem Controlling abhängig. Die Bewertungsgruppe spricht sich aufgrund der hohen Aussagekraft dieses Indikators dafür aus, die Erfassung von Drittmittelprojekten auf Ebene der Forschungseinheiten beizubehalten, auch weil die damit erfassten Projektbewilligungen eine eher in die Zukunft gerichtete Bewertung zulassen als die Verausgabungen. Um die Erhebung zu vereinfachen, sollte eine Bagatellgrenze in Höhe von 10.000 Euro in die Definition von Drittmittelprojekten eingeführt werden, so dass z. B. die Angabe von Reisemitteln und Kleinstprojekten in den Projektlisten der einzelnen Forschungseinheiten entfällt. Für Forschungseinheiten, die etwa idR kleinere Industrieprojekte bearbeiten, sollte eine summarische Aufstellung, etwa: „5 Industrieprojekte, Gesamtsumme 20.000 Euro“, möglich sein. Zusätzlich sollte für jede Forschungseinheit analog zu der bisher auf Einrichtungsebene abgefragten Tabelle eine nach Gebern und Jahren differenzierte Statistik der verausgabten Drittmittel erfragt werden. Nach den Erfahrungen aus der Pilotstudie wäre es dabei notwendig, mögliche Doppelmeldungen kontrollieren zu können; auf eine parallele Erfragung der verausgabten

Drittmittel auf Ebene der Einrichtung kann deshalb voraussichtlich nicht verzichtet werden.

C.II. Kriterium II, Impact/Effektivität[30]

Dem Kriterium Impact/Effektivität wurde von der Bewertungsgruppe ähnlich hohe Belastbarkeit bescheinigt wie der Forschungsqualität. Zu einzelnen Indikatoren sind folgende Hinweise zu machen:

- *Interdisziplinarität:* Interdisziplinarität war in der Operationalisierungsphase als ein eigenständiger Bewertungsaspekt diesem Kriterium zugeschlagen worden. Diese Entscheidung wurde in der Bewertungsphase revidiert, da zunehmend deutlich wurde, dass Interdisziplinarität kein eigenständiges Qualitätsmerkmal ist – weder ist Interdisziplinarität per se positiv, noch ist disziplinär orientierte Forschung geringer einzuschätzen. Der Versuch, das Ausmaß der Interdisziplinarität durch eine Zählung des Anteils derjenigen Zitationen zu messen, die nicht aus einem der chemischen Teilgebiete stammen, brachte aufgrund der Definitionen dieser Teilgebiete im WoS eher erratische Ergebnisse und sollte angesichts des Aufwands dieser Messung nicht wiederholt werden. Wegen dieser Probleme verzichtete die Bewertungsgruppe darauf, Interdisziplinarität als eigenständigen Bewertungsaspekt zu benoten. Der von den Einrichtungen vorgelegte Selbstbeschrieb zur Interdisziplinarität wurde aber als wertvolle Hintergrundinformation angesehen und half dabei, Forschungseinheiten an den Rändern der Chemie besser zu beurteilen.
- *Plenarvorträge:* Die hier gemachten Angaben waren sehr heterogen und zeugten häufig von dem Wunsch, durch eine möglichst lange Liste von Vorträgen eine positive Bewertung zu erzwingen. Aus diesem Grund wurden die definitorischen Vorgaben selten eingehalten und anstelle von Plenarvorträgen auf internationalen Kongressen jede Art von eingeladenen Vorträgen bis hin zu lokalen Seminaren eingetragen. Da es unmöglich war, tausende von Vorträgen auf den jeweiligen Tagungsprogrammen zu verorten, konnten diese Daten nicht adäquat bereinigt werden. Die Bewertungsgruppe schlägt vor, die Zahl der anzuführenden Plenarvorträge auf maximal fünf Angaben pro namentlich gemeldetem Wissenschaftler zu beschränken. Dabei muss klar gestellt werden, dass Angaben über diese Zahl hinaus nicht berücksichtigt werden und damit die Bewertung nicht positiv beeinflussen können. Die bislang ebenfalls unter „Vorträge“ erfassten „named lectures“

[30] Unter „Impact/Effektivität“ ist die Sichtbarkeit der Forschung in der scientific community zu verstehen.

könnten künftig zusammen mit Preisen, Akademiemitgliedschaften etc. als wissenschaftliche Auszeichnungen erfasst werden.

- Die Anzahl der angemeldeten und erteilten *Patente* wurde als zusätzliches Produktivitätsmaß neben der Publikationszahl bei diesem Kriterium mit herangezogen, spielte aber in erster Linie bei der Bewertung des Transfers eine Rolle. Für Einrichtungen mit einer stark transferorientierten Mission spielen Patente aber auch als Produktivitätsindikator eine Rolle, so dass sie trotz geringer Korrelation mit der Bewertung weiter als ergänzende Information verwendet werden sollten.
- Der *Anteil drittmittelfinanzierten Personals* ist als größenunabhängiges Maß der Drittmittelaktivität eher ein Effizienzindikator und sollte künftig diesem Kriterium zugeordnet werden.

Auffallend war in der gesamten Pilotstudie, dass die semantische Unterscheidung zwischen „Effektivität“ und „Effizienz“ schwierig war. Es wäre daher wünschenswert, eine andere Bezeichnung für das Kriterium „Impact/Effektivität“ zu finden. Die Bewertungsgruppe schlägt den Terminus „Sichtbarkeit“ vor.

C.III. Kriterium III, Effizienz

Bei der Bewertung der Effizienz führt die Definition des „Inputs“ über die Personalressourcen dazu, dass Unterschiede in der Ausstattung, die für die Produktivität des wissenschaftlichen Personals wichtig sind, nicht berücksichtigt werden konnten. Angesichts der großen Schwierigkeiten, unabhängig von der Art der Haushaltsführung und der Grundausstattung vergleichbare Informationen über die finanziellen Ressourcen von Einrichtungen unterschiedlichen Typs zu erhalten, ist dies dennoch der einzige mit vertretbarem Erhebungsaufwand gangbare Weg. Die Bewertungsgruppe Chemie ist der Ansicht, dass die Einzelfälle, in denen die Effizienzrelativierung auf den Faktor Personal weniger angemessen erschien, auf Basis der vorliegenden qualitativen Informationen durch die Berichterstatter und in der plenaren Diskussion angemessen berücksichtigt wurden. Die höhere Korrelation der Bewertungen mit den auf das grundfinanzierte Personal relativierten Indikatoren spiegelt die normative Überlegung wider, dass die Einwerbung von Drittmittelpersonal, das in den Nenner der auf das gesamte Personal relativierten Indikatoren eingeht, selbst ein Aspekt der Effizienzsteigerung ist. Dennoch sollten auch künftig beide Arten von Indikatoren berechnet werden.

Belastbare, auch einrichtungsübergreifend vergleichbare Zahlen zu den bei der Erbringung bestimmter Forschungsleistungen entstehenden Kosten zur Verfügung zu haben, ist auch unabhängig von Fragen der Forschungsbewertung ein großes politisches und volkswirtschaftliches Desiderat. Sobald die derzeitigen Bemühungen, im öffentlichen Sektor zu einem transparenten Rechenschaftswesen zu kommen, zu einer besseren Datenlage hinsichtlich der tatsächlichen Kosten geführt haben, sollten diese auch in die Effizienzbewertung eines künftigen Forschungsratings eingehen.

Ein zweiter kritischer Aspekt der Effizienzbewertung ist die Berücksichtigung der Lehrbelastung. Die für die Pilotstudie vorgenommenen pauschalen Relativierungen können lediglich die Unterschiede zwischen universitären und außeruniversitären Einrichtungen ausgleichen. Um zu einer genaueren Berechnung zu gelangen, müsste die tatsächliche Lehrbelastung erfasst werden. Angesichts der Verflechtung der Fächer durch Lehrimport und -export ist dies allerdings kaum durch eine einfache Kennzahl – etwa die Zahl der Hauptfachstudenten oder der abgenommenen Prüfungen – in fairer Weise möglich. Eine adäquate Erfassung würde den Erhebungsaufwand erheblich erhöhen und wäre vermutlich im Ergebnis kaum ausschlaggebend. Die Gutachter haben im Einzelfall Sondertatbestände berücksichtigt, etwa dass bei kleineren Einrichtungen eine hohe Lehrbelastung schwerer zu kompensieren ist als bei großen Einrichtungen.

Der zentrale Indikator der Zitationen pro VZÄ (wissenschaftliches Personal) ist, anders als der zentrale Qualitätsindikator, nicht auf Teilgebiete normiert. Ein statistischer Zusammenhang der Effizienzbewertungen mit der Spezialisierung der Einrichtungen auf bestimmte Fachgebiete, die anhand der durchschnittlichen FCS_m-Werte ihrer Publikationen bestimmt werden kann, ist nicht nachweisbar. Dennoch wäre wünschenswert, künftig auch die Effizienzbewertung fachspezifisch normieren zu können. Die Option, dafür einen fachnormierten absoluten Zitationswert Z/FCS_m heranzuziehen, sollte gemeinsam mit Experten der bibliometrischen Forschung geprüft werden.

C.IV. Kriterium IV, Nachwuchsförderung

Die zur Bewertung der Nachwuchsförderung ausgewählten Indikatoren liegen schwerpunktmäßig im Bereich der akademischen Nachwuchsförderung. Dies macht die faire Bewertung von Einrichtungen, die vor allem Nachwuchsförderung für den außerakademischen Bereich betreiben, schwierig. Es wäre wünschenswert, wenn

auch Universitäten längerfristige Absolventenverbleibsstatistiken führen würden, um künftig auch die Erfolge der Nachwuchsförderung im außerakademischen Bereich erfassen zu können.

Anmerkungen zu einzelnen Indikatoren:

- *Promotionszahlen* sind größenabhängig und spiegeln somit die Bedeutung der Einrichtung für die Nachwuchsförderung aus Sicht des Gesamtsystems wider. Verschiedentlich wurde vorgeschlagen, die Promotionen auf die Zahl der Professoren zu relativieren. Die so berechneten, größenunabhängigen Werte wären ein Maß der Effizienz der Nachwuchsförderung; sie sollten nicht mit den Outputindikatoren konfundiert werden.
- Größere Bedeutung für die Bewertung als die Promotionszahlen haben die *Stipendienzahlen* und die Zahl der *Stellen für Nachwuchswissenschaftler* (Doktorandenstellen, Nachwuchsgruppenleiter). Hier gab es jedoch relativ häufig Definitions- oder Erfassungsprobleme. Die Qualität dieser Daten kann möglicherweise durch eine dezentrale Erhebung bei den einzelnen Forschungseinheiten verbessert werden, wie sie im Fach Soziologie vorgenommen wurde.
- *Promotionen außeruniversitärer Einrichtungen* werden bei den gewählten Erhebungsmodalitäten doppelt gezählt. Außeruniversitär betreute Promotionen bei den jeweiligen Universitäten herauszurechnen hat aber auf die relativen Positionen der Universitäten keinen signifikanten Einfluss.
- Der Indikator *Anteil weiblicher Promovenden* korreliert nicht mit der Bewertung. Da die Daten von der GDCh ohnehin erhoben werden, besteht unter gleichstellungspolitischen Aspekten keine Notwendigkeit, sie für ein Forschungsrating gesondert auszuwerten.

C.V. Kriterium V, Transfer in andere gesellschaftliche Bereiche

Die zur Bewertung des Transfers in andere gesellschaftliche Bereiche ausgewählten Indikatoren erfassen fast ausschließlich den Transfer in die Wirtschaft. Dies ist für das Fach Chemie ein angemessener Weg, Erfolge in der Anwendung zu erfassen, und spiegelt einen wichtigen Aspekt der Leistung der chemischen Forschung wider. Bei einer Wiederholung des Verfahrens in der Chemie sollte dieses Kriterium deshalb in „Transfer in die Wirtschaft“ umbenannt werden. Für andere Fächer sind andere, dem für sie typischen Weg des Transfers angemessene Indikatoren zu entwickeln.

Zu den einzelnen Indikatoren sind folgende Anmerkungen zu machen:

- In der Erhebung wurde versucht, *Drittmittel von Unternehmen* von Mitteln für *Auftragsforschung* abzugrenzen. Diese Abgrenzung konnte von vielen Einrichtungen inhaltlich nicht nachvollzogen werden, nur ein Drittel von ihnen machte überhaupt Angaben zur Auftragsforschung. Wo die Unterscheidung bekannt war, wurden Mittel für Auftragsforschung zum Teil nicht zentral erfasst, da diese Mittelflüsse über Konten der einzelnen Professuren abgewickelt wurden. Wo Auftragsforschungsmittel angegeben wurde, machten diese nur wenige Prozent der Drittmittel von Unternehmen aus und scheinen mit diesen im übrigen gut zu korrelieren. Auf die gesonderte Erfassung von Auftragsforschung sollte deshalb künftig verzichtet werden.
- *Patente und Lizenzen* sind ein wichtiger Indikator für Transfer in die Anwendung. Da Patenttitel auf Ebene der Forschungseinheiten gesondert erfasst wurden, konnten zum Teil erhebliche Diskrepanzen zwischen den dezentralen Angaben und den zentralen Patentstatistiken nachgewiesen werden. Dies liegt zum Teil daran, dass Patente aus Kooperationsprojekten mit Unternehmen häufig von diesen und nicht von den Hochschulen angemeldet werden. Es bleibt abzuwarten, ob die Änderung des Arbeitnehmererfindungsgesetzes, die in den Erhebungszeitraum fiel, mittelfristig zu einer Verbesserung der Patenterfassung durch die Hochschulen führt. Andernfalls ist zu prüfen, ob analog zur Publikationsrecherche auch eine unabhängige Patentrecherche durch ein darauf spezialisiertes Institut in Auftrag zu geben ist.
- *Ausgründungen/Spin-Offs* sind ein wichtiger, aber eher qualitativ zu wertender Indikator. Im internationalen Vergleich sind die Zahlen in Deutschland relativ gering, was möglicherweise damit zu tun hat, dass die chemische Industrie in Deutschland sehr gut aufgestellt ist, so dass die Umsetzung von Erfindungen in der Regel in Kooperationsprojekten mit bereits bestehenden Unternehmen einfacher zu bewerkstelligen ist als durch Unternehmensneugründungen.
- Ein interessanter zusätzlicher Indikator könnte die Zahl der Stellen sein, die nach einem definierten Zeitraum durch einen Spin-Off generiert wurden. Dies ist nach momentaner Datenlage bei den Einrichtungen indes nicht zu erfassen; zudem ist eine aussagekräftige Zahl vermutlich erst nach einer Frist zu bestimmen, die für ein Forschungsrating deutlich zu lang ist.
- Ein wichtiger Transferindikator sind *Beratungsleistungen*. Diese unterliegen jedoch häufig Geheimhaltungsbestimmungen. Anonymisierte Angaben können naturge-

mäß nicht überprüft werden. In der Chemie, wo öffentliche Auftraggeber, die keine derartigen Geheimhaltungsauflagen machen, eine vergleichsweise geringe Rolle spielen, ist deshalb überlegen, ob auf diese Daten verzichtet werden kann.

C.VI. Kriterium VI, Wissensvermittlung und -verbreitung

Für das Kriterium Wissensvermittlung und -verbreitung wurde nur ein einziger quantitativer Indikator verwendet, die Zahl der abgeschlossenen Berufsausbildungen in chemienahen Berufen. Im übrigen beruhte die Bewertung ausschließlich auf qualitativen Angaben, die sehr heterogen waren. Vor allem die Angaben zu Projekten der Wissensvermittlung (*Weiterbildungskurse, Transferveranstaltungen*) waren sehr unterschiedlich stark detailliert. So wurden teils Einzelveranstaltungen von zweistündiger Dauer einzeln aufgelistet, teils mehrjährige Programme pauschal geschildert. Bei einer erneuten Erhebung sollten hier Wege gesucht werden, eine stärkere Standardisierung der Angaben zu erzielen. Es könnte etwa pauschal nach einzelnen Kategorien gefragt werden, z.B. ob 1. Weiterbildungskurse, 2. Einzelveranstaltungen und 3. Schülerlabors o.ä. durchgeführt wurden, inklusive einer summarischen Auflistung der Teilnehmerzahlen und exemplarischen Erläuterungen zu den Veranstaltungen. Es wäre zudem denkbar, dass die Einrichtungen als Nachweis ihrer Vermittlungsarbeit bis zu 10 Zeitungsartikel, Transkripte von Rundfunksendungen etc. einreichen könnten.

Da das Engagement in der Wissensvermittlung und -verbreitung zu den Aufgaben der Hochschulen gehört und für die Rekrutierung des wissenschaftlichen Nachwuchses wie auch das gesellschaftliche Ansehen des Faches von Bedeutung ist, wäre für das Fach Chemie ein Verzicht auf Kriterium VI keine Option. Vielmehr sollten die Unterschiede ziwschen beiden Kriterien durch eine Spezifizierung von Kriterium V als „Transfer in die Wirtschaft“ und Kriterium VI als „Wissensvermittlung in die Öffentlichkeit“ deutlicher gemacht werden.

Die Option, die Kriterien V und VI getrennt beizubehalten, ist nicht ohne weiteres auf andere Fächer zu übertragen. Die Dimension Wissenstransfer wurde in den Empfehlungen des Wissenschaftsrates zu Rankings im Wissenschaftssystem als eine stark fachabhängige Leistungsdimension gesehen;[31] dies hat sich in der Bewertung des Faches Chemie bestätigt.

---

[31] Wissenschaftsrat: Empfehlungen zu Rankings im Wissenschaftssystem. Teil 1: Forschung. in: Empfehlungen und Stellungnahmen 2004, Köln 2005, S. 205.

Anlage: Ergebnisse der Pilotstudie Forschungsrating Chemie