Stefan Hornbostel, Anna Schelling (Hg.)

EVALUATION: NEW BALANCE OF POWER?

iFQ-Working Paper No. 9 | November 2011

Institut für
Forschungsinformation
und Qualitätssicherung

iFQ – Institut für Forschungsinformation und Qualitätssicherung

Schützenstraße 6a
10117 Berlin

Telefon 030-206 41 77-0
Fax 030-206 41 77-99
E-Mail info@forschungsinfo.de
Internet www.forschungsinfo.de
www.research-information.de

ISSN 1864-2799

November 2011

Stefan Hornbostel, Anna Schelling (Hg.)

EVALUATION: NEW BALANCE OF POWER?

iFQ-Working Paper No. 9 | November 2011

Evaluationen sind aus dem deutschen Wissenschaftssystem nicht mehr wegzudenken: Evaluiert werden Forschungsförderprogramme, Hochschulen, Forschungseinrichtungen und einzelne Fächer. Insbesondere seit Evaluationen zunehmend als Entscheidungsgrundlage für Ressourcenzuweisungen genutzt werden, stehen sie immer stärker im Fokus von Politik und Öffentlichkeit und sehen sie sich häufig Kritik ausgesetzt.

Die 4. iFQ-Jahrestagung, die wir gemeinsam mit der Forschungsgruppe Wissenschaftspolitik des Wissenschaftszentrums Berlin für Sozialforschung durchgeführt haben, wollte Gelegenheit bieten, darüber nachzudenken, wie derartige Beurteilungsprozesse – nicht nur in der Wissenschaft – ablaufen, wie sie organisiert und wie reflektiert werden:

Was ist Evaluation?

Wer sind die Evaluatoren, wie kommen sie zu ihrem Status und wie gehen sie damit um?

Wie laufen Evaluationen in der Praxis ab?

Welche Auswirkungen haben Evaluationen auf die Evaluierten?

Wie verändern Evaluationen Wissenschaftspolitik und Wissenschaftssteuerung?

Diesen Fragen widmeten wir uns auf der iFQ- Jahrestagung 2010 aus verschiedenen Perspektiven, um so zu einem besseren Verständnis aktueller Evaluationspraktiken sowie ihrer Wirkungen und Begleiterscheinungen beizutragen. So interessierten das Wesen von Evaluationen und die Rolle der Gutachter – der Evaluierenden – in der Welt von Wissenschaft und Forschung und darüber hinaus. Aus Sicht der Ritualforschung näherte sich die Tagung der Evaluation als modernem Ritual. Positive und negative Erfahrungen mit Evaluationen, Begleiterscheinungen und Folgen von Evaluationen wurden aus der Sicht der Evaluierenden und der Evaluierten betrachtet.

Das vorliegende iFQ-Working Paper dokumentiert die Ergebnisse der Jahrestagung 2010 des iFQ zum Thema „Evaluation: New Balance of Power?“. Die Tagung wurde vom iFQ gemeinsam mit der Forschungsgruppe Wissenschaftspolitik am Wissenschaftszentrum Berlin für Sozialforschung durchgeführt.

Inhalt

Stefan Hornbostel

Resonanzkatastrophen, Eigenschwingungen, harmonische und chaotische Bewegungen

Am 7. November 1940, kurz nach Fertigstellung der in Rekordzeit gebauten Tacoma Bridge (Spannweite 843 m) im US-Bundesstaat Washington, versetzte ein kräftiger Seitenwind die Brücke in Schwingungen. Bei dieser Brücke nichts Ungewöhnliches (sie hatte den Spitznamen „Galloping Gertie"), aber an diesem Tag wurden die Schwingungen immer größer, führten zu immer heftigeren Verwindungen des Fahrdecks und schließlich zum Einsturz der Brücke. Die durch den Wind ausgelöste „selbsterregte Schwingung" war mit immer weiterer Energie versorgt worden, bis es schließlich zur Katastrophe kam. Sechzig Jahre später war das Schwingungsproblem noch keineswegs gelöst. Die Millennium Bridge – eine Fußgängerbrücke über die Themse in London – musste nur zwei Tage nach ihrer Eröffnung im Juni 2000 wegen unkontrollierter Schwingungen für den Publikumsverkehr gesperrt werden. Diesmal war es nicht der Wind, sondern es waren die Fußgänger, die kleine, zufällige Schwingungen durch reflexhafte Ausgleichbewegungen verstärkten. Unglücklicherweise trafen sie damit die Eigenfrequenz der Brücke, so dass sich eine Resonanzschwingung ergab. Da Menschen bei allzu starkem Schwanken dazu neigen stehenzubleiben und sich festzuhalten, blieb die Resonanzkatastrophe aus. Denn ohne weitere Energiezufuhr kommt ein schwingendes System aufgrund der Dämpfung nach und nach zum Stillstand.

Derartige Rückkopplungen kommen nicht nur in technischen Systemen vor, sondern auch in biologischen, geologischen, klimatischen, wirtschaftlichen und sozialen Systemen. Ob es dabei zu immer stärkeren Aufschaukelungen oder zu einer Abschwächung kommt, hängt insbesondere davon ab, ob eine positive oder negative Rückkopplung stattfindet. Handelt es sich um ein Regelsystem, wird abwechselnd der eine oder der andere Prozess genutzt, um das System im gewünschten Gleichgewichtszustand zu halten.

In der Biologie sind vielfältige derartige Regelungsprozesse bekannt (z.B. Osmose, cirkadiane Rhythmen), in Ökosystemen überlagern sich in komplexer Weise die Wachstums- und Schrumpfungsraten von Räuber- und Beutepopulationen, aus der Klimaforschung sind die Wechselwirkungen zwischen Vereisung und Sonnenlichtreflektion bekannt. In sozialen Netzwerken gewinnen Nachrichten erst dann an Bedeutung, wenn sie Resonanz ausgelöst haben und andere Teilnehmer sich der Kommunikation anschließen; aus der Wirtschaft sind nicht nur die langen Kondratjew-Wellen bekannt, sondern auch das Auf und Ab der Börsenkurse, die entstehen, weil alle Akteure – völlig regelkonform – dasselbe tun. Entfernt man die Dämpfung (im letzteren Fall die Verantwortlichkeit für das eigene Handeln in Gestalt des wirtschaftlichen Bankrotts im Falle des Scheiterns, Stichwort: „too big to fail") aus einem solchen System, tritt die Resonanzkatastrophe früher oder später ein, so wie wir sie jüngst in der Finanzkrise und ihren anhaltenden Folgen erlebt haben und noch erleben. Diese Krise zeigt auch, wie gut gemeinte Dämpfungsmaßnahmen in ihr Gegenteil umschlagen und auf unerwartete Weise das System erneut anregen können.

In technischen Systemen hilft der nachträgliche Einbau einer Dämpfung in aller Regel. So wurde die Millenium Bridge mit 58 Schwingungsdämpfern aufgerüstet und nach zwei Jahren Arbeit für den Fußgängerverkehr wieder freigegeben. In sozialen Systemen ist eine solche Sperrung und Nachrüstung hingegen nur schwer oder gar nicht möglich. Die Anregung oder Dämpfung von schwingungsfähigen sozialen Systemen ist also eine durchaus riskante Angelegenheit, zumindest dann, wenn die Anregungsfrequenz mit der Eigenfrequenz übereinstimmt. Diese eher technische Vorstellung prägt auch die soziologische Systemtheorie. Luhmann (1986:40) begreift Resonanz als „rekursiv-geschlossene Reproduktion bei umweltoffener Irritierbarkeit". Autopoietische Systeme schließen ihre interne Reproduktion gegen die Umwelt ab und werden nur ausnahmsweise durch Resonanz in Schwingung versetzt. Das System kann danach nur aufgrund seiner Eigenfrequenz zu

resonanten Schwingungen veranlasst werden, was heißt, dass die Anregung in Gestalt der für das jeweilige Subsystem typischen binären Codierung wirken muss. Auch wenn man eine weniger strenge Theoriearchitektur anlegt und derartige Systeme als Akteurskonfigurationen betrachtet, die nicht in autopoietischer Abgeschlossenheit operieren, sondern lediglich im Rahmen eines spezialisierten Sinnhorizonts (vgl. Mayntz 1988), bleibt aus Sicht der Systemtheorie, dass die „Realität" eines anderen Subsystems nur in Form von Beobachtungen, genauer durch die Beobachtung von Beobachtungen, zugänglich ist, und das macht zumindest die planvolle Anregung eines sozialen Systems schwierig (vgl. Taubert (erscheint 2011)).

Zweifellos kennt auch das Forschungssystem zyklische Bewegungen (Innovationszyklen, Expansions- und Kontraktionsbewegungen, Forschungsmoden etc.). Die Anregungen dieses Systems sind teils endogener Natur, entstammen also den in der Forschung selbst aufgeworfenen Fragen und Problemen, teils exogener Natur in Gestalt von ökonomischen, technischen und sozialen Herausforderungen, die insbesondere in Form von wettbewerblichen Förderprogrammen (mit einer peer review-gestützten Auswahl) in das Wissenschaftssystem „übersetzt" werden.

Derartige exogene Stimulationen erfolgten in der Nachkriegszeit in Deutschland zunächst mit Fokus auf die Lehre und führten zur Expansion des Hochschul- und Forschungssystems und gleichzeitig zur Differenzierung in Universitäten und Fachhochschulen in den 1960er und 70er Jahren. Der schnellen Expansion folgte eine Stagnationsphase. Als Dämpfungsfaktoren wurden dabei die unterschiedlichsten Phänomene ausgemacht: von allgemeiner finanzieller Unterausstattung und Reformstau über unangemessene staatliche Detailsteuerung, unvollständige Organisationsformen (die im Verein mit einem hohen Maß an Selbststeuerungskompetenzen nolens volens zu einem Stillstandspakt führten) und mangelnde vertikale Differenzierung des Systems bis hin zu fehlenden Steuerungs- und Anreizsystemen reichten die Diagnosen. Einen späten Ausdruck fand diese Stillstandsituation (nicht nur in der Wissenschaft) 1997 in der sogenannten „Ruck-Rede" des Bundespräsidenten, in der er dafür eintrat, dass ein Ruck durch Deutschland gehen müsse, um die „verkrusteten Strukturen" zu überwinden. Allerdings waren zu Beginn der 1990er Jahre bereits erste „Oszillatoren" in Gestalt von Rankings und Evaluationen installiert, die für eine noch zaghafte erste „Anregung" des Hochschul- und Forschungssystems sorgten. In der Folge nahmen Art und Menge derartiger „Anregungen" wie auch die Quellen der Impulse sehr schnell zu: Diverse Reformen der rechtlichen Rahmenbedingungen, Änderungen des Besoldungssystems, nationale und internationale Rankings und Ratings, leistungsorientierte Mittelverteilungssysteme, routinemäßige Evaluationen, Genderprogramme, Globalhaushalte, z.T. weitreichende Erhöhungen der Hochschulautonomie, große Förderprogramme von Bund und Ländern (darunter die Exzellenzinitiative), Förderinitiativen von Stiftungen, Veränderungen in der Drittmittelförderung, gezielte Internationalisierungen, Versuche, die Kooperationen zwischen Universitäten und außeruniversitären Forschungseinrichtungen zu stimulieren, erhöhte mediale Aufmerksamkeit für die Forschungs- und Bildungspolitik und vieles mehr.

Terminologisch wurden diese höchst unterschiedlichen Maßnahmen unter dem Begriff „New Public Management" versammelt, mit dem zumindest die basalen Trends bezeichnet wurden, wie etwa der Rückzug des Staates aus der Detailsteuerung, unternehmerisches und strategisches Handeln der Wissenschaftsorganisationen, am Output orientierte Incentives, die Verstärkung markt- und wettbewerbsförmiger Elemente und die Koexistenz sehr unterschiedlicher Steuerungs- und Integrationsmodi.

Auch wenn das Konzept des „New Public Management" gegenüber hergebrachten Begriffen wie „Kontextsteuerung" sehr unbestimmt bleibt, wird doch ein wesentliches Element deutlich: das Fehlen zentraler Steuerungsinstanzen und damit auch das Fehlen eines politischen „Masterplans". Vom Zurückdrängen der politischen Steuerungs- und Kontrollansprüche wurde die Mobilisierung „unternehmerischer" Aktivitäten und insgesamt eine Effizienzsteigerung erwartet. In Kauf genommen wurde dafür, dass sich nicht nur die intendierten Effekte, sondern auch die nicht

intendierten Wirkungen (einschließlich der kaum kalkulierbaren Interferenzen unterschiedlicher Maßnahmen) nach der Logik von Märkten entwickelten.

Dieser Wechsel vom klassischen „Government" zur „Governance" ist allerdings empirisch nur schwer zu fassen: einerseits, weil kausale Zurechnungen in einem derartigen Wirkungsgeflecht nur schwer möglich sind, andererseits, weil es sich um graduelle Veränderungen handelt und mithin sehr unterschiedliche Integrations- und Steuerungsmodi nebeneinander wirken.

Drei Beispiele sollen im Folgenden die potentiellen Wirkungen und Wechselwirkungen von externen Anregungen auf das Schwingungsverhalten sozialer Systeme illustrieren:

I Eingeübte Evaluationen: unerwartete Vorzüge des Rituals

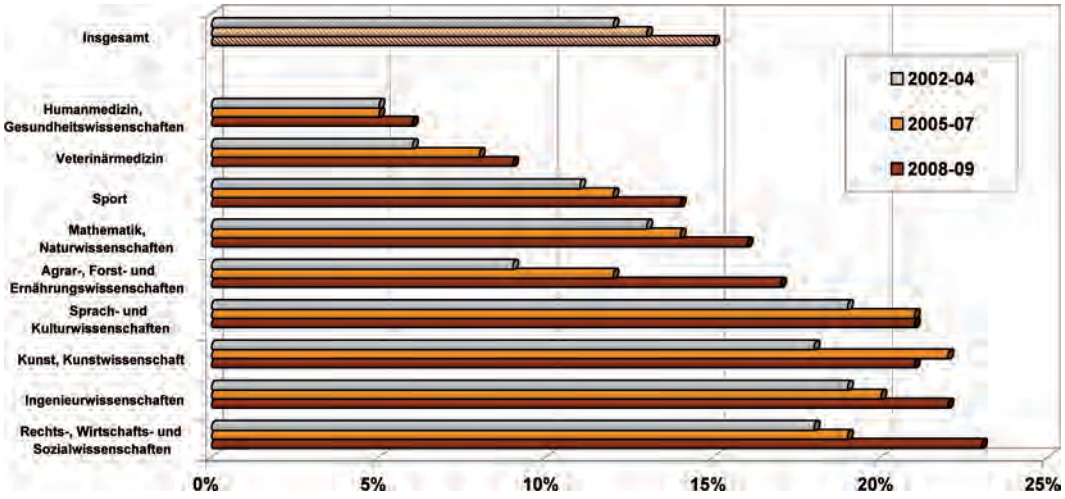
Lehr- und Forschungsevaluationen gehören seit fast zwei Jahrzehnten zu den „Oszillatoren", die, nicht nur über Transparenzerzeugung, sondern zum Teil auch durch handfeste unmittelbare Folgen, Bewegung im Wissenschaftssystem erzeugen. Wenngleich von derartigen Evaluationen häufig wichtige und für die Einrichtungen äußerst positive Effekte ausgegangen sind, provoziert ihre Wiederholung – in manchen Einrichtungen nach einem genau festgelegten Prozedere – strategische Reaktionen auf Seiten der Evaluierten. Das Spektrum reicht von der mechanistischen Umsetzung von Empfehlungen bis hin zu ausgedehnten Probeläufen, in denen eine Selbstinszenierung im Hinblick auf die antizipierten Gutachtererwartungen oder aber mit Bezug auf eigenständig definierte – dem disziplinären Wertekanon entnommene – Leistungsdimensionen entworfen wird, in Rollenspielen die Begutachtungssituation simuliert wird, das Personal gebrieft, trainiert und gegebenenfalls zum Termin beurlaubt wird und nach Analyse der Stärken und Schwächen ein Plan zur Invisibilisierung der Schwachstellen entworfen wird. Das alles scheint auf den ersten Blick eine sehr ressourcenintensive Inszenierung zu sein, in der – eine gute Regie vorausgesetzt – nur noch erfahrene Evaluatoren, die allerdings ihrerseits immer stärker als Gutachter nachgefragt werden, in der Lage sind, einen Blick hinter die Kulissen oder die Fassaden eines potemkinschen Dorfes zu werfen. Abgesehen davon, dass gute Regisseure selten sind und auch Dramatisierungen und Inszenierungen einen guten Ausgangsstoff benötigen, zeigt sich auf den zweiten Blick noch etwas ganz anderes: Ethnologen und Theologen wissen um die heilsame Wirkung von Ritualen. Sie liegt nicht darin, dass der vorgebliche Zweck der Ritualhandlung erreicht wird, sondern in der Stiftung von Orientierung und Gruppenidentität und in der symbolischen Auseinandersetzung mit grundlegenden Fragen, die im Alltagsgeschäft kaum thematisierungsfähig sind. Rituale können, aber müssen keineswegs sinnentleerte Routinen sein; sie bieten vielmehr einen nicht begründungsbedürftigen Anlass für Reflexion und Motivation. Insofern dürften von Evaluationen jenseits der mehr oder weniger gelungenen Prüfung von Zielerreichungen erhebliche Impulse ausgehen, die in ihrer Wirkung aber nur schwer prognostizierbar sind.

II Incentives und Inflation: Rankingfolgen

Fast jedes Ranking (aber auch die meisten leistungsorientierten Mittelverteilungssysteme) listet unter den präsentierten Forschungsindikatoren auch die Anzahl der Promotionen auf (als absolute Zahl oder bezogen auf die Zahl der Professoren). Dabei wird viel und gut in Eins gesetzt. Es geht nicht um die Qualität der Promotionen, die Karriere der Doktoranden oder gar um den Forschungsertrag der Dissertation, sondern schlicht um die Menge der abgeschlossenen Promotionen. Da auch unabhängig von Rankings die Nachwuchsförderung hoch im Kurs steht, ist ein Wettbewerb um Doktoranden entstanden. Aus einer marktwirtschaftlichen Sicht (mit ceteris paribus-Bedingungen) sollte das Ergebnis erfreulich sein: Die besten Doktoranden kommen an die besten Ausbildungsstätten und die werden wiederum mit steigender Reputation belohnt. Wie die Geschichte lehrt (vgl. Hornbostel 2009) und einige prominente Politiker demonstriert haben, gibt es allerdings auch andere Wege zur Promotion, nämlich durch Senkung der Qualitätsstandards. Den Universitäten verschafft dies Wettbewerbsvorteile, die Promovierenden sparen – sofern sie keine Wissenschaftskarriere im Auge

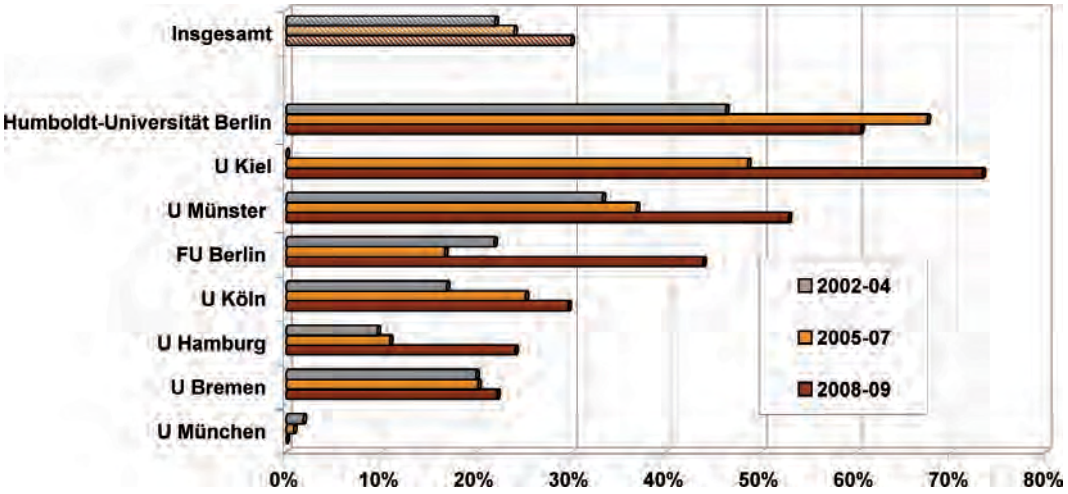
haben – Zeit und Geld. Ein Nachweis für eine derartige Tendenz ist schwer zu erbringen, zumal alle offiziellen Stellungnahmen beteuern, dass die Qualität der deutschen Promotion unzweifelhaft sei, auch wenn einzelne „schwarze Schafe" einen anderen Eindruck erwecken und der European Research Council die deutsche Medizinerpromotion nicht als Qualifikationsnachweis für eine Bewerbung um einen „Starting Grant" anerkennen will. Dennoch stimmen zwei Umstände nachdenklich. Erstens werden deutsche Promovenden beständig besser (vgl. Abb.1). Vom Zeitraum 2002 - 2004 bis zum Zeitraum 2008/2009 stieg der Anteil der exzellenten Promotionen (Note: summa cum laude) bundesweit von 12 Prozent auf 15 Prozent.

Abbildung 1: Promotionen – Anteile „summa cum laude", nach Fächergruppen, 2002-2009



Quelle: DeStatis 2011, eigene Berechnungen

Abbildung 2: Wirtschaftswissenschaft: Anteil der „summa cum laude"-Promotionen an allen Promotionen



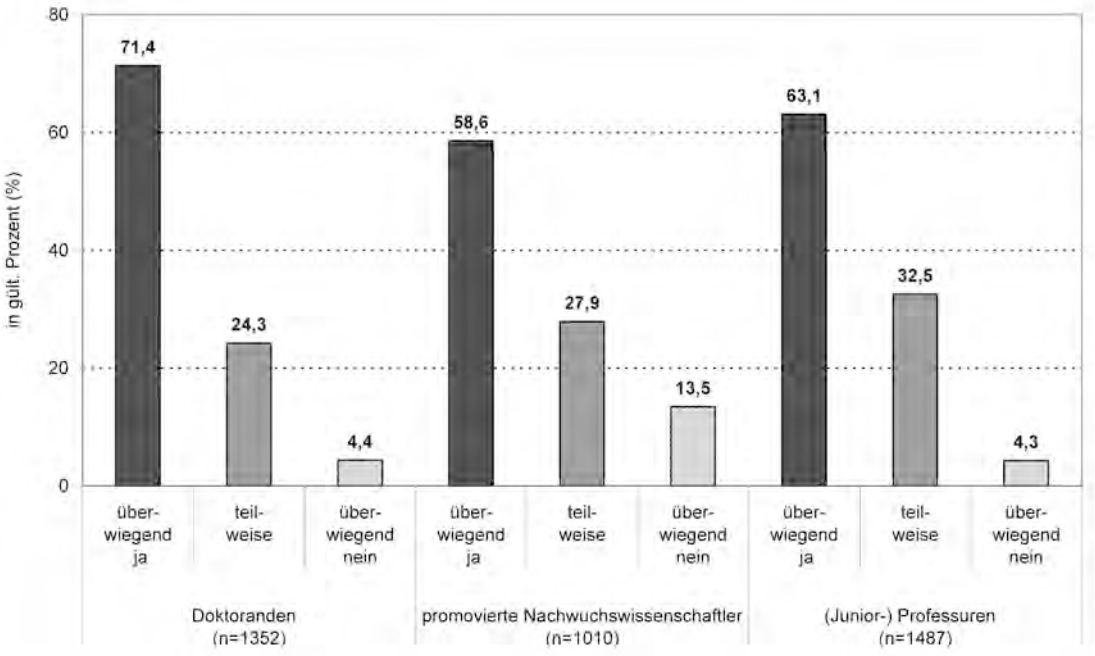
Quelle: DeStatis 2011, eigene Berechnungen

Beeindruckender sind die Steigerungen in einzelnen Fächergruppen. So konnten die Rechts-, Wirtschafts- und Sozialwissenschaften sich von 18 Prozent auf 23 Prozent steigern. Und noch weitaus beeindruckender sind die Steigerungsraten einzelner Fakultäten und die Differenzen zwischen den Universitäten. Die Universität Kiel konnte etwa im genannten Zeitraum den Anteil der „summa

cum laude"-Promotionen im Fachgebiet Wirtschaftswissenschaften von 0 auf 73 Prozent steigern, während die Universität München (LMU) von knapp 2 Prozent auf 0 zurückfiel (vgl. Abb. 2).

Ganz offenkundig entsprechen die Verhältnisse nicht den ceteris paribus-Annahmen der reinen Marktlehre. Wichtiger aber ist zweitens, dass die scheinbar an einzelnen Orten im Überfluss produzierte Qualität nicht mit den Erfahrungen der Professoren bei der Personalrekrutierung übereinstimmt, denn in der Wissenschaftler-Befragung des iFQ (Böhmer et al. 2011) gaben die Professoren gerade bei der Rekrutierung von promovierten Postdocs die größten Schwierigkeiten an (vgl. Abb. 3). Der wesentlichste Grund, der für die Rekrutierungsschwierigkeiten angegeben wurde, war ein Mangel an geeigneten Kandidaten.

Abbildung 3: Wissenschaftler-Befragung 2010 des iFQ: „Konnten für die zu besetzenden Positionen Personen mit den gewünschten Qualifikationsprofilen gewonnen werden?"



Quelle: Böhmer et al. 2011: Wissenschaftler-Befragung 2010: Forschungsbedingungen von Professorinnen und Professoren an deutschen Universitäten. iFQ-Working Paper No.8. Bonn.

Es scheint also trotz aller offiziellen Beteuerungen im Zusammenspiel von Rankings, leistungsorientierter Mittelvergabe und den Bemühungen um eine Steigerung der Doktorandenzahlen zu einer inflationären Abwärtsspirale sowohl im Hinblick auf die Qualifikationsprofile der Doktoranden gekommen zu sein als auch im Hinblick auf die Begutachtungs- und Benotungspraxis.

III Anregung und Dämpfung: Die „gimme five"-Regelung der DFG

Die Forschungsförderung gibt ein anschauliches Beispiel für das Wechselspiel von Anregung und Dämpfung. Im Jahr 1997 legte die Deutsche Forschungsgemeinschaft (DFG 1997) erstmals ihr fortan regelmäßig erscheinendes Förderranking vor und schloss sich damit einer seit Ende der 1980er Jahre bestehenden Form der Beobachtung von Wissenschaft durch ein Set von Wissenschaftsindikatoren an, konkret durch Berichterstattung über die Drittmitteleinnahmen der Hochschulen. Diese quantitativen Indikatoren sind nicht nur ein Beobachtungsinstrument, sie wirken vielmehr über leistungsorientierte Mittelverteilungssysteme und durch die mediale Öffentlichkeit unmittelbar

auf die Verteilung von wissenschaftlicher Reputation, institutionelle Strategien und individuelle Akquisemotivationen. Immerhin gaben 63,5 Prozent der Professoren und Professorinnen in der iFQ Wissenschaftler-Befragung (Böhmer et al. 2011) an, dass ihre Grundausstattung für Forschung zumindest zum Teil von Drittmitteleinwerbungen abhinge.

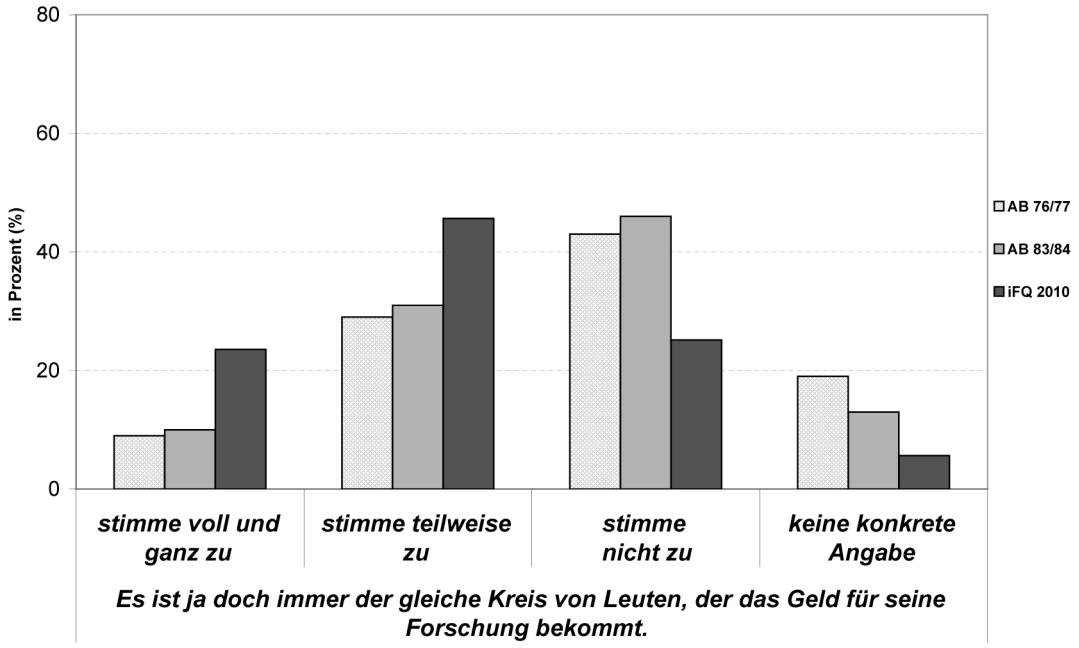
Im Jahre 2010 wird aber genau diese Art der Beobachtung mit ihren Implikationen für das Verhalten von Wissenschaftlerinnen und Wissenschaftlern problematisiert. Die im Februar 2010 bekannt gegebenen Veränderungen in den Antragsregularien der DFG sehen vor, dass „statt beliebig vieler Veröffentlichungen nur noch wenige, besonders aussagekräftige Publikationen als Referenz" genannt werden dürfen. Begründet wird das Vorgehen damit, dass die „immer größere Bedeutung von […] numerischen Indikatoren verringert werden" soll. „'Ob bei der leistungsorientierten Mittelvergabe, bei Habilitationen und Berufungen und auch bei den Bewertungen von Förderanträgen – überall haben numerische Indikatoren wie der Hirsch-Faktor oder der Impact-Faktor immer mehr Gewicht bekommen. […] Das übt einen außerordentlich starken Druck auf Wissenschaftlerinnen und Wissenschaftler aus, möglichst viel zu publizieren. Und es verleitet immer wieder zu Fällen wissenschaftlichen Fehlverhaltens, in denen falsche Angaben zum Stand einer Veröffentlichung gemacht werden. Das alles schadet der Wissenschaft', betonte der DFG-Präsident" (DFG 2010).

Während einerseits also das System stimuliert wird und numerische Größen wie das Volumen der eingeworbenen Drittmittel offeriert und häufig durch leistungsorientierte Mittelverteilungssysteme der Bundesländer und Universitäten zur Ressourcenallokation genutzt und damit in der Bedeutung weiter verstärkt werden, wird auf der anderen (output-bezogenen) Seite eine Dämpfung eingezogen durch die Begrenzung der Publikationsliste. Ungeachtet einiger praktischer Schwierigkeiten mit der neuen Regelung (Fachkollegiaten und Gutachter möchten in der Regel eine Überblick über das gesamte Œuvre haben, Befangenheitsprüfungen setzen ebenfalls die gesamte Publikationsliste voraus, Publikationskulturen sind sehr unterschiedlich, so dass eine fixe Begrenzung fachspezifisch sehr unterschiedlich wirkt), könnte diese Regelung potentiell das Gegenteil des intendierten Effekts erzielen. Zumindest kann dieser Eindruck entstehen, wenn man die italienischen Erfahrungen mit einer ähnlichen Regelung im Rahmen der nationalen Evaluation berücksichtigt. Dort ließ sich feststellen, dass die Universitäten keineswegs jene Publikationen benannten, die auf den bibliometrischen Indikatoren besonders gut abschneiden, sondern Publikationen auswählten, die nach internen Kriterien (einschließlich den der Macht- und Reputationshierarchie geschuldeten) als berücksichtigungswürdig galten (vgl. Abramo/D'Angelo/Caprasecca 2009). Ähnliches ließ sich in Deutschland auch beim Rating des Wissenschaftsrates feststellen. Die Folgen eines solchen Auseinanderklaffens von internen Auswahlkriterien und bibliometrischen Indikatoren sind absehbar: Da die Bedeutung der bibliometrischen Indikatoren nicht abnimmt, wenn nur eine begrenzte Zahl von Publikationen zur Beurteilung einer Person oder einer Organisation zugelassen wird, sondern im Gegenteil eher zunimmt – denn nun sind Gutachter natürlich daran interessiert, etwas über die ausgewählten Publikationen zu erfahren – entsteht ein strategisches Interesse daran, jene Publikationen auszuwählen, die auf den bibliometrischen Indizees gut abschneiden (quod erat impedendum). Zumindest in Italien richten sich entsprechend einige Universitäten darauf ein, bei der Auswahl der Publikationen für das nationale Evaluationsverfahren mit bibliometrischer Beratung behilflich zu sein. Nicht auszuschließen, dass auch bei der Formulierung von Drittmittelanträgen demnächst die Bibliometrie das letzte Wort bei der Auswahl der zu benennenden Publikationen hat. Eine zweifellos gut gemeinte Maßnahme kann auf diese Weise leicht eine Resonanzschwingung erzeugen, obwohl eine Dämpfung intendiert war.

Der partielle Wechsel vom „Government" zur „Governance" hat zweifellos Bewegung in das Wissenschaftssystem gebracht und dabei Verkrustungen unterschiedlicher Art aufgebrochen – die Dämpfung des Systems deutlich verringert. Zugleich aber hat die Vielzahl der Reformen und Programme, die sich teils verstärken, teils neutralisieren, eine zunehmend kritische Haltung hinsichtlich der Kosten und der Verteilungswirkung einer immer stärker wettbewerblich

organisierten Forschung hervorgebracht. Deutlich erkennbar wird diese Veränderung an einem Survey Item, das schon Mitte der 1970er Jahre im Rahmen der Allensbacher Hochschullehrerbefragung deutschen Professoren vorgelegt wurde (vgl. Abb. 4).

Abbildung 4: Wissenschaftler-Befragung 2010 des iFQ: „Es ist ja doch immer der gleiche Kreis von Leuten, der das Geld für seine Forschung bekommt." (N=2.714)



Legende: AB 76/77: Allensbacher Hochschullehrerbefragung 1976/77, AB 83/84: Allensbacher Hochschullehrerbefragung 1983/84, iFQ 2010: iFQ Wissenschaftler-Befragung 2010

Quelle: Böhmer et al. 2011: Wissenschaftler-Befragung 2010: Forschungsbedingungen von Professorinnen und Professoren an deutschen Universitäten. iFQ-Working Paper No.8. Bonn.

Die kräftig gestiegene Überzeugung unter den Professoren aller Fachgebiete, dass die Forschungsmittel sich stark konzentrierten und die Verteilung nicht nur nach meritokratischen Kriterien erfolge, bildet ihrerseits eine „Anregung" des Systems, die sich auf das Verhalten von Wissenschaftlern auswirkt. Im schlimmsten Fall als Missachtung der Rahmenbedingungen fairen Wettbewerbs, nämlich den Regeln guter wissenschaftlicher Praxis (vgl. Martinson/Anderson/de Vries 2005, Martinson et al. 2006).

Bleibt also zu hoffen, dass dem deutschen Wissenschaftssystem das Schicksal von „Galloping Gertie" erspart bleibt und ein guter Mix von Dämpfung und Anregung zu Stabilität führt, wie bei der Millenium Bridge.

Literatur

Abramo, Giovanni / D'Angelo, Ciriaco Andrea / Caprasecca, Alessandro, 2009: Allocative efficiency in public research funding: Can bibliometrics help? Research Policy 38, 206-215.

Böhmer, Susan / Neufeld, Jörg / Hinze, Sybille / Klode, Christian / Hornbostel, Stefan, 2011: Wissenschaftler-Befragung 2010: Forschungsbedingungen von Professorinnen und Professoren an deutschen Universitäten. iFQ-Working Paper No.8. Bonn.

Deutsche Forschungsgemeinschaft, 1997: Bewilligungen nach Hochschulen. Bewilligungsvolumen 1991 bis 1995, Bonn.

Deutsche Forschungsgemeinschaft, 2000: DFG-Pressemitteilung Nr. 7 | 23. Februar 2010 http://www.dfg.de/service/presse/pressemitteilungen/2010/pressemitteilung_nr_07/index.html

Hornbostel, Stefan, 2009: Promotion im Umbruch – Bologna ante Portas, in: *Martin Held, Gisela Kubon-Gilke, Richard Sturn (Hg.):* Jahrbuch Normative und institutionelle Grundfragen der Ökonomik, Band 8, Bildungsökonomie in der Wissensgesellschaft. Marburg: Metropolis Verlag, 213-240.

Luhmann, Niklas, 1986: Ökologische Kommunikation. Kann die moderne Gesellschaft sich auf ökologische Gefährdung einstellen? Opladen: Westdeutscher Verlag.

Martinson, Brian C. / Anderson, Melissa S. / de Vries, Raymond, 2005: Scientists behaving badly. Nature 435 (9 June 2005), 737-738.

Martinson, Brian C. / Anderson, Melissa S. / Crain, A. Lauren / de Vries, Raymond, 2006: Scientists' Perception of Organizational Justice and Self-Reported Misbehaviors. Journal of Empirical Research on Human Research Ethics, 51-66.

Mayntz, Renate, 1988: Funktionelle Teilsysteme in der Theorie sozialer Differenzierung, in: *Mayntz, Renate / Rosewitz, Bernd / Schimank, Uwe / Stichweh, Rudolf (Hg.):* Differenzierung und Verselbständigung. Zur Entwicklung gesellschaftlicher Teilsysteme. Frankfurt/M.: Campus, 11-44.

Taubert, Niels, 2011 (im Erscheinen): Bibliometrie in der Forschungsevaluation. Zur Konstitution und Funktionslogik wechselseitiger Beobachtung zwischen Wissenschaft und Politik, in: *Wehner, Josef / Passoth, Jan (Hg.),* Web 3.0. Wiesbaden: VS Verlag.

Wilhelm Krull

Bewertung, Begutachtung und Evaluation in Wissenschaft und Forschung

Im Laufe der letzten beiden Jahrzehnte ist „Evaluation" zu einem schillernden Begriff geworden. Es scheint so, als ob jede Art von Leistungsmessung, Begutachtung und begleitender Bewertung bereits der Anhauch von Evaluation umgibt. Manche Kritiker sprechen deshalb auch bereits von einer „Evaluitis" (Wolfgang Frühwald), die nahezu alle Bereiche des wissenschaftlichen Lebens erfasst habe.

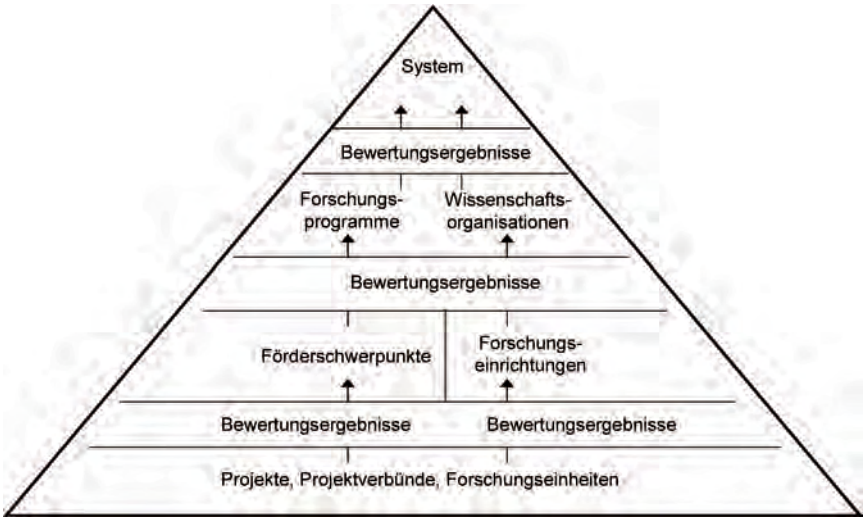
Vor diesem Hintergrund erscheint es sinnvoll, zunächst einmal die Begrifflichkeiten zu klären, dann einen Blick auf die vermessene Wissenschaftswelt zu werfen und schließlich die Entwicklung der Evaluationen im deutschen Wissenschaftssystem rückblickend zu bewerten.

1 Was heißt und zu welchem Ende betreibt man Evaluationen?

Die hiermit aufgeworfene Frage lehnt sich eng an Friedrich Schillers bedeutsamen Aufsatz mit dem Titel „Was heißt und zu welchem Ende betreibt man Universalgeschichte?" an. Schiller hat seinerzeit in seiner Antrittsvorlesung an der Universität Jena klar unterschieden zwischen dem „Brotgelehrten", der keinerlei akademische Ambitionen verfolgt, sondern lediglich in dem Professorenberuf die Chance sieht, einen angenehmen Lebensunterhalt zu gewährleisten, und dem „philosophischen Kopf", der sich mit hohem Engagement dem Erkenntnisfortschritt verschreibt und diesen wiederum auch an seine Studierenden weitergibt. Nun mag mancher von Ihnen denken, dass diese Unterscheidung nun doch schon mehr als 200 Jahre alt sei und deshalb kaum noch Gültigkeit beanspruchen könne. Wenn man freilich bedenkt, dass auch heute mehr als die Hälfte der Professoren und Professorinnen an deutschen Hochschulen niemals einen Antrag bei der Deutschen Forschungsgemeinschaft stellt, dann liegt zumindest die Vermutung nahe, dass auch in unserer Zeit die „Brotgelehrten" nicht ganz ausgestorben zu sein scheinen. Dieser Teil der Professorenschaft entzieht sich also weitgehend dem Wettbewerb um Drittmittel und wird somit von international vergleichender Antrags- und Projektbegutachtung nicht erfasst. Erst die externe Evaluation der jeweiligen Organisationseinheit könnte daher die mangelnden Forschungsaktivitäten einzelner Personen aufdecken. Die Frage bleibt freilich, inwieweit mittels Evaluation dieser Umstand tatsächlich aufgegriffen und verändert werden kann? Oder ob nicht doch bloß mittels immer mehr und immer weiter ausgreifender Evaluationen „rituals of verification" (Michael Power: The Audit Society, 1997) vollzogen werden.

Um den Dschungel der Begrifflichkeiten und Rituale ein wenig zu lichten, erscheint es zunächst einmal notwendig, klar zu unterscheiden zwischen Bewertung, Begutachtung und Evaluation. Während Bewertung einen konstitutiven Bestandteil des wissenschaftlichen Lebens darstellt und sich allenthalben sowohl in Studium und Lehre als auch in Wissenschaft und Forschung vollzieht, setzt Begutachtung bereits voraus, dass eine dritte Instanz jemanden um eine Stellungnahme, z.B. im Zuge von Berufungsverfahren, Drittmittelvergaben oder Zeitschriftenpublikationen, bittet. Evaluation hingegen ist auf einer höher aggregierten Ebene angesiedelt. Evaluiert werden z.B. Forschungsprogramme, Institutionen oder ganze Wissenschaftssysteme. Evaluationen erfordern also demnach einen Vorlauf an begutachteten und bewerteten Vorhaben, Personen und/oder Forschergruppen. Größere Zusammenhänge können nur adäquat evaluiert werden, wenn entsprechend der nachfolgend dargestellten Evaluationspyramide ein umfassender Informationsfluss von unten nach oben gewährleistet ist.

Abbildung 1: Die Evaluationspyramide



Eine Evaluation sollte niemals ohne konkreten Anlass und klare „Terms of Reference" in Auftrag gegeben werden, damit sie die notwendigen Informations-, Qualitätssicherungs- und gegebenenfalls Allokationsfunktionen erfüllen kann. Für die Evaluation von Forschungsinstituten ist letztlich entscheidend, dass das Zusammenspiel von einer ex ante-Evaluation der Konzepte und Leitungspersönlichkeiten mit der begleitenden Bewertung der Institutsarbeit durch Fachbeiräte und den alle fünf bis sieben Jahre stattfindenden institutsübergreifenden, vergleichenden Evaluationen sorgfältig abgestuft und zueinander in Beziehung gesetzt wird. Mit Blick auf die Evaluation von Forschungs- und Entwicklungsprogrammen sind ferner die folgenden fünf Komponenten entscheidend:

Evaluation von F&E Programmen

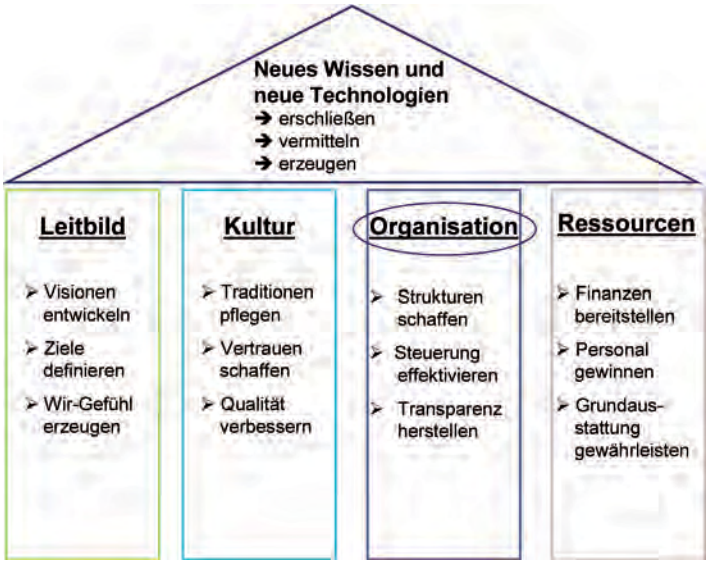
1. **Ziele, Zeiten, Zusammenarbeit**
2. **Organisatorischer Rahmen**
   - Unabhängigkeit und Transparenz
3. **Methoden und Techniken**
   - Wissenschaftlicher Erfolg: Peer Review, Bibliometrie etc.
   - Sozioökonomische Effekte: Datenanalyse, Befragungen
   - Implementierung und Management: Prozessanalysen, Interviews
   - breit angelegte Wirkungsanalysen: Studien
4. **Berichtssystem**
   - Präsentation der Ergebnisse
   - Rückmeldung, Interaktion
5. **Umsetzung der Ergebnisse**
   - Prozessorientierung.

Seit Mitte der 1980er Jahre haben wir in einer Reihe von europäischen Ländern die Einführung von Einrichtungen zur Evaluation der Hochschulen beobachten können. Neben dem Higher Education Funding Council in Großbritannien und dem Comité National d'Évaluation in Frankreich sowie einer von den Hochschulen selbst eingesetzten Instanz in den Niederlanden haben wir im Laufe der Zeit auch mit dem Schweizer Wissenschaftsrat, dem Evaluierungszentrum in Dänemark sowie der Wissenschaftlichen Kommission des Landes Niedersachsen weitere Ausformungen solcher Institutionen gesehen. Die besondere Problematik, die sich im Unterschied zu in der Regel gut ausgestatteten außeruniversitären Forschungseinrichtungen im Bereich der Hochschulen ergibt, liegt

vor allem darin, dass wir es in letzterem Fall mit einem kranken System zu tun haben, in dem die Grundvoraussetzungen für eine Evaluation der Lehr- und Forschungsleistungen nur bedingt gegeben sind. Statt mit institutioneller Eigenverantwortung haben wir es nach wie vor in weiten Teilen mit Politikverflechtung zu tun. Die Unterfinanzierung der Hochschulen und die insbesondere in den großen Fächern zu beklagenden schlechten Betreuungsrelationen liegen so auf der Hand, dass eine Kritik an diesen Zuständen nicht eigens eine Evaluation erfordert, sondern zunächst einmal der Abhilfe bedarf.

Zugleich muss klar sein, dass eine Universität nur dann von externen Evaluationen profitieren kann, wenn sie sich als autonome und zugleich als lernende Organisation begreift. Dies verdeutlicht die nachfolgende Grafik, indem sie die zentralen Säulen der modernen Forschungsuniversität – Leitbild, Kultur, Organisation und Ressourcen – beleuchtet, welche von Relevanz sind für die Wahrnehmung der zentralen Aufgaben unserer Hochschulen: neues Wissen und neue Technologien zu erschließen, die Kenntnisse darüber zu vermitteln und vor allem auch teilzuhaben an der Erzeugung neuen Wissens.

Abbildung 2: Die Universität als lernende Organisation



Die damit verbundene Neuformierung der Universitäten erscheint umso dringlicher angesichts der hohen Veränderungsdynamik, der sich die Hochschulen in ihrem Umfeld ausgesetzt sehen. Wenige Stichworte mögen hier genügen, um die Veränderungsrichtung anzudeuten:

- von der Information zur Informatisierung des Wissens,
- von der Interdisziplinarität zur problemorientierten, transdisziplinären Forschung,
- von der bi- und trilateralen Internationalisierung zur Globalisierung,
- von der öffentlichen zur privaten Finanzierung,
- von der inputorientierten Hochschulplanung zur outputorientierten Leistungsbewertung.

Evaluation kann nur erfolgreich sein, wenn sie in ein gut durchdachtes Kommunikationskonzept eingebettet wird und Handlungsfolgen erzeugt. Die Voraussetzungen für solch eine erfolgreiche Evaluation sind vor allem:

- Klare Definition der Ziele der Evaluation
- Verständigung über geeignete Kriterien, Methoden, Techniken und Vergleichsmaßstäbe

- Gemeinsames Verständnis über Chancen und Nutzen der Evaluation
- Intensive Vorbereitung der Evaluation
- Transparenter Umgang mit Informationen
- Offenheit gegenüber den Evaluatoren und Interesse an deren Hinweisen
- Intensive und vorurteilsfreie Diskussion der Empfehlungen und ihrer Umsetzung.

2 Die vermessene Wissenschaftswelt

Ratings und Rankings sind in den letzten beiden Jahrzehnten wie Pilze aus dem Boden geschossen. Allenthalben werden die verschiedensten Zusammenstellungen von Indikatoren genommen, um eine neue Rangliste präsentieren zu können. Auch wenn viele dieser quantitativen Bewertungen den Anspruch erheben, seriös und fundiert zu sein, tut man gleichwohl als Wissenschaftspolitiker oder Mitglied einer Hochschulleitung gut daran, auf diese Ranglisten nicht zu bauen, wenn es gilt, strategische Weichenstellungen vorzunehmen. Man sollte sie vielmehr als Teil der Unterhaltungsindustrie betrachten und sich von allzu kurzatmigen Schlussfolgerungen fernhalten, wie dies auch indirekt Christoph Schneider in einem Artikel nahegelegt hat: „So wie Midas im griechischen Mythos alles zu Gold werden ließ, was er berührte, und darüber verhungerte, so wird dem Ranglisten versessenen Evaluator alles zur Zahl, die ihm bald die Wirklichkeit verstellt." (Frankfurter Allgemeine Zeitung vom 1.10.2009).

Wenn wir gleichwohl einen Blick auf das derzeit bekannteste Ranking, nämlich das „Shanghai Ranking" (2009) der Jiaotong-Universität werfen, dann müssen wir in der Tat feststellen, dass Europa weder unter den Top 5 noch den Top 100 oder auch 200 bemerkenswert stark vertreten ist. Freilich ändert sich das Bild, wenn wir die Top 500 Universitäten der Welt betrachten, unter denen Europa immerhin mit deutlich mehr Universitäten vertreten ist als die USA.

Abbildung 3: Globaler Wettbewerb – das Shanghai Ranking 2009

	Amerika	Europa	Asien/Pazifik	Afrika
Top 5	4	1	-	-
Top 20	17	2	1	-
Top 50	39	10	2	-
Top 100	59	32	9	-
Top 200	99	79	22	
Top 300	134	125	42	1
Top 400	162	170	67	2
Top 500	184	208	106	3

Dies spiegelt zugleich den Weg wider, den wir in Europa in nahezu allen Ländern seit den 1960er Jahren gegangen sind. Dieser Weg implizierte, dass wir Hochschulentwicklung zugleich als Teil der Regionalentwicklung begreifen wollten, um auf diese Weise wissenschaftsintensive Forschung und Entwicklung möglichst breit zu streuen. Für internationale Spitzenforschung gab es dabei wenig Raum.

Trotz dieser auf den ersten Blick einleuchtenden Bewertung der rund 500 Universitäten darf man freilich nicht verkennen, dass die Kriterien, die dem Shanghai-Ranking zugrunde liegen, einen klaren Bias in Richtung Natur- und Technikwissenschaften aufweisen, der von den europäischen

Universitäten, vor allem auch von den traditionell breit in den Geistes- und Gesellschaftswissenschaften aufgestellten Hochschulen, keineswegs adäquat abgebildet werden kann. Die sechs wichtigsten Kriterien und ihre Gewichtungen sind wie folgt angelegt:

- Die Zahl der Alumni, die einen **Nobelpreis** in den Fächern Physik, Chemie, Medizin oder Wirtschaftswissenschaften sowie eine Fields-Medaille in der Mathematik erhalten haben (10 Prozent).
- Die Zahl der Fakultätsangehörigen, die einen Nobelpreis oder eine Fields-Medaille gewonnen haben (20 Prozent).
- Die Zahl der Artikel mit einer Koautorschaft von Fakultätsmitgliedern, die in **Nature und Science** publiziert wurden (20 Prozent).
- Die Zahl der Artikel, die von Fakultätsmitgliedern der Universität publiziert wurden und im **Science Citation Index** oder im **Social Science Citation Index** gelistet sind (20 Prozent).
- Die **Zahl der vielfach zitierten Forscher** der Universität in 21 breit angelegten Themenfeldern (20 Prozent).
- Die akademische Leistung in Relation zur Größe der Universität (10 Prozent).

Einen Überblick über die derzeit gängigsten nationalen und internationalen Rankings geben die folgenden Übersichten:

Abbildung 4: Nationale Rankings

Name des Rankings	Spezialisierung	Aktualität
Der Spiegel		1999/2007
DFG-Förderranking	DFG-Bewilligungen	2009
Focus		2007
Handelsblatt	BWL und VWL	2009/2010
Hochschulanzeiger (FAZ)	Wirtschaftshochschulen	2006
Humboldt-Ranking	Forschung / Internationale Attraktivität	2009
Karriere		2009
Mein Prof	Kursbewertungen	2009
Studi-VZ		2009
Wirtschaftswoche Forscherranking		Einzelstudie
Wirtschaftswoche Uniranking		2009

Abbildung 5: Internationale Rankings

Name des Rankings	Spezialisierung	Aktualität
Financial Times Global MBA-Ranking	MBA	2010
HEEACT	Wissenschaftliche Arbeiten / Bibliometrie	2009
Leiden-Ranking	Bibliometrie	2009
QS Topuniversities		2010
Shanghai Ranking (ARWU)		2009
Times Higher Education World Universities Rankings		2009
US News		2009
Webometrics	Webseiten / Internetperformance	2010

Quelle: www.che-ranking.de

Neben den bereits angesprochenen Imbalancen in der Berücksichtigung der verschiedenen Fächerkulturen (insbesondere der Vernachlässigung der Geistes- und Gesellschaftswissenschaften) muss ferner bei der Betrachtung solcher Rankings berücksichtigt werden, dass sie nahezu alle auf den bibliometrischen Daten in den von Thomson Scientific lieferbaren Datenbeständen beruhen.

Letztlich können solche Bewertungen nur funktionieren in Fächern, für die nicht nur die zitierenden, sondern auch die weit überwiegende Mehrheit der zitierten Veröffentlichungen in der Datenbank enthalten sind. Hier gibt es jedoch ebenfalls große Diskrepanzen zwischen den Fächergruppen. Die Kongruenz liegt bei

- fast 100 % in den Naturwissenschaften,
- 40 bis 60% in der Mathematik und in den Wirtschaftswissenschaften,
- unter 15% in den Geisteswissenschaften.

Es kommt hinzu, dass Zitationen als Indikatoren von Forschungsqualität eine Reihe gravierender technischer Probleme aufweisen, die nicht zuletzt interdisziplinäre Arbeiten negativ betreffen. Grundsätzlich handelt es sich jedoch um fundamentale Fehlerquellen wie z.B.:

- Die Erfassung der Daten ist gespickt mit Fehlern.
- Das Organisationsprinzip des *citation index* sind Fachgebiete – interdisziplinär arbeitende Forscher sind im Nachteil.
- Grund für die häufige Zitation eines Aufsatzes muss nicht dessen Qualität, sondern kann auch dessen Fehlerhaftigkeit sein.
- Der *citation index* rechnet jedes Zitat jedem einzelnen Autor eines Artikels zu – die Einzelleistung wird nicht sichtbar / honoriert.

Insbesondere in solchen Wissenschaftssystemen, die die Publikationserfolge in referierten Zeitschriften und vor allem die Impactfaktoren solcher Zeitschriften zur direkten Leistungsbewertung und -belohnung zugrunde gelegt haben, können wir seit einigen Jahren beobachten, dass dies die Betriebsamkeit enorm erhöht. Insoweit ist Alfred Kieser zuzustimmen, der bereits 2003 festgestellt hat: „Wie alle Evaluationskriterien schafft auch der citation index die Wirklichkeit, die zu messen er vorgibt. Er ändert das Verhalten der Evaluierten." (in: Die Zeit, vom 17.07.2003). Wir können dies insbesondere auch in Deutschland darin beobachten, dass es einen klaren Trend hin zu mehr Aufsätzen statt Monographien gibt, dass interessantes Material zur Erhöhung des Impactfaktors „verlängert" wird und die Kumulation von Artikeln bereits ausreicht, um zu promovieren oder sich zu habilitieren. Der Konservatismus mit Impact-Garantie dominiert gegenüber risikobehafteter, transformativer Forschung in neuen Gebieten, auf denen zunächst nur eine kleine Community die Arbeiten zitieren kann. Dies hat zur Folge, dass Standardisierung vorherrscht und die Produktion von Aufsätzen mit potenziell hohem Impactfaktor im Vordergrund steht gegenüber problemadäquater interdisziplinärer Forschung.

In der vermessenen Hochschulwelt scheinen Wissenschaftspolitiker wie Hochschulleitungen zu vergessen, dass man

- die Drittmittelquote eines Historikers nicht mit der eines Ingenieurs,
- die Kosten für einen Absolventen an einer Forschungsuniversität nicht mit denen für einen Absolventen an einer Fach- oder Regionalhochschule,
- die Anzahl der Promotionen pro Professor/Professorin an einer medizinischen nicht mit denen an einer philosophischen Fakultät vergleichen kann und sollte.

3 Evaluation – ein Blick zurück und nach vorn

Obwohl es bereits in den 1980er Jahren Evaluationen des Wissenschaftsrates von Blaue Liste-Instituten gab, muss doch von heute her gesehen das Jahr 1990 als die entscheidende Zäsur im deutschen Wissenschaftssystem gesehen werden. Mit den Empfehlungen des Wissenschaftsrates zu den „Perspektiven für Wissenschaft und Forschung auf dem Weg zur Deutschen Einheit" (vom

Juli 1990) und der anschließenden Evaluation der Akademie-Institute der DDR wurde ein neues Kapitel aufgeschlagen, dem anschließend sich auch das westdeutsche Wissenschaftssystem nicht entziehen konnte. Zunächst folgten nach der Wiedervereinigung 1992/93 erste forschungsfeldbezogene Evaluationen durch den Wissenschaftsrat (z.B. zu den Umweltwissenschaften und den Materialwissenschaften). Während diese noch Ausfluss von Evaluationen der ostdeutschen Forschungsinstitute waren, ergab sich Ende der 90er Jahre eine völlig veränderte Situation dadurch, dass Bund und Länder die Systemevaluation aller großen Wissenschaftsorganisationen beschlossen. Dabei erfolgte zunächst die Evaluation der Fraunhofer-Gesellschaft durch eine weitgehend national besetzte Gutachterkommission. 1998/99 wurde dann die Systemevaluation der Deutschen Forschungsgemeinschaft (DFG) und der Max-Planck-Gesellschaft (MPG) durch eine internationale Kommission, die ich gemeinsam mit zwei Mitarbeitern betreut habe, initiiert. 2000 folgte der Bericht des Wissenschaftsrates zur Systemevaluation der Blauen Liste sowie 2001 ebenfalls seitens des Wissenschaftsrates der Bericht zur Systemevaluation der Helmholtz-Gemeinschaft Deutscher Forschungszentren.

Es ist von vornherein klar gewesen, dass eine solche Systemevaluation ganz andere Zeithorizonte in den Blick nehmen muss als etwa die Evaluation eines einzelnen Forschungsinstituts. Dementsprechend hat die internationale Kommission zur Evaluation von DFG und MPG die Entwicklung beider Organisationen in den letzten drei Jahrzehnten sowie die Anforderungen an ein künftiges Wissenschaftssystem näher untersucht. Dabei war es im Hinblick auf die Bewertung der Aktivitäten der DFG wichtig, die Überlastungssymptome in den Universitäten adäquat zu berücksichtigen. Ebenso wichtig war es freilich, die Interaktion zwischen der Max-Planck-Gesellschaft und den Universitäten sowie das Geflecht von Geldgebern und autonomen Wissenschaftsorganisationen näher zu betrachten.

Der Bericht hat von vornherein den konkreten Umsetzungskontext der Empfehlungen in den Vordergrund gestellt und Synopsen der wichtigsten Empfehlungen für Bund und Länder ebenso wie die beiden Wissenschaftsorganisationen und die Universitäten markiert. Von heute her gesehen kann man sicherlich mit Recht behaupten, dass viele Veränderungen in der Deutschen Forschungsgemeinschaft (z.B. die Öffnung für neue Förderformen bis hin zur Exzellenzinitiative) und bei der MPG (insbesondere das starke Zugehen auf die Universitäten mit Blick auf Max-Planck-Research Schools und die Beteiligung an Clustern und Graduiertenschulen in der Exzellenzinitiative) ohne die seinerzeitigen Impulse nicht in dieser Weise denkbar gewesen wären.

Nun wird freilich mancher von Ihnen fragen, „Was passiert denn eigentlich im Stiftungsbereich?“. In der Tat ist auch auf diesem Feld einiges in Gang gekommen. Nicht zuletzt maßgeblich dafür war die Arbeit an den „Grundsätzen guter Stiftungspraxis“ im Rahmen des Bundesverbandes Deutscher Stiftungen. Darin findet sich u.a. folgende Passage: „Die Stiftungsorgane sorgen für die regelmäßige Überprüfung der Wirksamkeit der Stiftungsprogramme vor allem im Hinblick auf die Verwirklichung des Satzungszwecks, die Effizienz des Mitteleinsatzes und im Hinblick auf das Verhalten gegenüber Fördersuchenden sowie der Öffentlichkeit; sie fördern entsprechendes Verhalten ihrer Mitarbeiter.“

Die VolkswagenStiftung hat mit Blick auf ihr 50jähriges Bestehen im Jahre 2012 bereits 2005 damit begonnen, ihre verschiedenen Förderinitiativen und -bereiche einer externen Evaluation zu unterziehen. Dabei ist zu berücksichtigen, dass es in einer Stiftung keinen externen Auftraggeber geben kann. Als autarke und autonome Institution kann nur durch das Aufsichtsgremium (also im Falle der VolkswagenStiftung durch das Kuratorium) der Auftrag an externe Experten ergehen, die entsprechenden Bereiche zu evaluieren. Dabei ist – wie eingangs bereits betont – auch für die VolkswagenStiftung essenziell, dass die Bewertung von Förderprojekten und -initiativen bereits vorliegt, wenn die Evaluationskommission für die Gesamtbetrachtung der Leistungen der VolkswagenStiftung ihre Arbeit aufnimmt. Dies wird durch die folgende Evaluationspyramide illustriert:

Abbildung 6: Evaluationspyramide der VolkswagenStiftung

4 Fazit

Angesichts der hohen Veränderungsdynamik stellen sich für jede langfristig angelegte Evaluation heutzutage neue Probleme darin, die Bewertung vergangener Leistungen zugleich zum Maßstab dafür zu machen, Empfehlungen für die Zukunft zu entwickeln. Daher ist es essenziell, soweit wie möglich zu versuchen, Evaluation und Prospektion miteinander zu verknüpfen. Zugleich ist es unbedingt erforderlich, von Anfang an die Folgen und die Implementationsaspekte stärker in den Blick zu nehmen. Ohne ein Monitoring des Umsetzungsprozesses bleiben in der Regel viele Empfehlungen auf dem Papier zurück. Ein weiteres wichtiges Problem, dass es mit Blick auf die Zukunft von Evaluationen anzugehen gilt, ist die vielfach gegebene Inkonsistenz zwischen den Evaluationsverfahren auf den unterschiedlichen Ebenen und die tatsächliche Durchführung von Wirkungsstudien. Hier ist noch viel Neuland zu betreten.

Der Trend zu immer mehr Evaluationen birgt zugleich vielerlei Gefahren:

- Folgenlose Evaluationen verkommen zu einem sinnentleerten Ritual
- Die „Evaluationsbürokratie“ lähmt den Wissenschaftsbetrieb
- Indikatorengläubigkeit verformt und verhindert kreative Forschung
- Falsch gewählte Benchmarks vermindern die Aussagekraft von Evaluationen
- Ein unter „Dauerevaluation“ stehendes Forschungsinstitut, -gebiet oder -system beschäftigt sich nur noch mit sich selbst statt mit seinen eigentlichen Aufgaben.

Ein chinesisches Sprichwort besagt: „Wenn der Wind des Wandels weht, bauen die einen Mauern, die anderen Windmühlen.“ Um nicht den Effekt des „sich einmauerns“ zu erzeugen, ist es wichtig, sich über den Wirkungshorizont von Evaluationen ein klares Bild zu verschaffen. Vor allem ist dabei zu bedenken, dass keine mid-term- oder ex post-Evaluation die Fehler ausbügeln kann, die im Bereich der ex ante-Bewertung von Forschungsleistungen oder auch von zu berufenden Wissenschaftlerinnen und Wissenschaftlern gemacht wurden. Insgesamt gesehen komme ich zu dem Schluss, dass

- Evaluation ein wirksames, aber kein Allheilmittel ist zur Verbesserung der Qualität von Wissenschaft und Forschung;
- dieses Mittel auf der Ebene ganzer Fächer oder Institutionen sparsam (ca. alle 5-7 Jahre)

und auf der Ebene großer Wissenschaftsorganisationen oder gar des ganzen Systems noch sparsamer (ca. alle 10 bis 15 Jahre) eingesetzt werden muss;

- Evaluationen eine kontinuierliche Qualitätskontrolle nicht ersetzen können;
- Evaluationen stets anlass- und strukturbezogen in Auftrag gegeben werden sollten;
- eine erfolgreiche Evaluation einer ausführlichen Vorbereitung bedarf:
  - für die Entwicklung von Erfolgskriterien und Benchmarks
  - für das Sammeln der notwendigen Informationen
  - für das Vermitteln der jeweiligen Vorgehensweisen;
- eine Evaluation nicht mit dem Vorstellen des Evaluationsberichts, sondern mit der Umsetzung ihrer Ergebnisse endet.

Axel Michaels

Evaluation als akademisches Ritual[1]

Ich beginne mit einem Bekenntnis: Ich bin gegen Evaluationen in der Wissenschaft. Jedenfalls bin ich gegen zu viele Evaluationen. Ich spreche dabei aus der Opferperspektive, als jemand, auf dessen (virtuellem) Schreibtisch jeden zweiten Tag etwas zum Evaluieren liegt und der sich morgens überlegen muss, ob er nach links zum Stapel der Akten, die ein Gutachten erfordern, oder nach rechts zu eigenem Forschungsmaterial greift. Ich schätze, dass ich im Jahr etwa 200-250 Anträge zu begutachten habe. Das macht – Wochenenden als Arbeitstage kalkulatorisch, wenn auch nicht faktisch ausgenommen – etwa eine Begutachtung pro Tag.

Ich teile in diesem Dilemma zwischen Forschung und Forschungsbeurteilung die sattsam bekannte Gründe gegen Evaluationen, wie sie etwa der Schweizer Wirtschaftswissenschaftler Bruno S. Frey vorbringt (Frey 2007):

- Evaluationen werden in ihrer Wirkung überschätzt. Das Ziel, Exzellenz zu fördern, wird trotz Evaluationen keineswegs befriedigend erreicht.
- Das Messbare (Drittmittel, Publikationen, Auszeichnungen und Preise, Zitationen) steht im Vordergrund.
- Es gibt falsche Belohnungen: Die Wissenschaftsmanager[2] werden gegenüber den einzelnen Forschern bevorzugt.
- Die kurzen Fristen und Zeiträume der Evaluationen sind den Forschungen mitunter nicht angemessen: Manchmal braucht ein Wissenschaftler ein ganzes Leben, um den großen Wurf hinzubekommen.
- Radikale, disziplinär abweichende Ideen werden bisweilen behindert, da der Anpassungsdruck an die herrschenden Meinungen groß ist.
- Es gibt einen vorauseilenden Gehorsam, die Evaluationskriterien zu erfüllen – mit der Folge, dass Ergebnisse hochgejubelt oder gar gefälscht werden.
- Die Kosten, auch und gerade die versteckten oder indirekten Kosten, sind zu hoch und werden unterschätzt.
- Misstrauen statt Vertrauen wird gesät: Ein im harten Ausleseprozess ausgewählter Forscher verliert seinen Vertrauensvorschuss, also das, was er am meisten braucht.

An dieser Stelle muss ich noch etwas gestehen: Ich habe meinen Beitrag vorab einer Evaluation unterziehen lassen. Eigene Gedanken und eigenes Wissen sind gut, aber Kontrolle ist besser! Man kennt das. Der Evaluator bringt folgendes Monitum an (nicht weil es nötig gewesen wäre, sondern weil es erwartet wird – wie es meiner Erinnerung nach bei H.C. Artmann mit Bezug auf die erste Fußnote eines von ihm selbst herausgegebenen eigenen Manuskripts heißt): „Die obigen Darlegungen zeugen von einer voreingenommenen und daher unwissenschaftlichen Haltung. Die Behauptungen sind nicht auf dem Stand der Forschungsliteratur oder empirisch belegt. Der Referent versteht nichts von der Audit Society *eines Michael Power (1997), zitiert nicht die relevante Literatur, weder Christine Schwarz' Buch über* Evaluation als modernes Ritual (Schwarz 2006)*, noch Robert Floden und Stephen Weiners Aufsatz* „Rationality to Ritual" (Floden/Weiner 1978)*, und auch nicht die vielen einschlägigen Artikel*

1 Eine kürzere Version dieses Beitrags erscheint in der Zeitschrift Gegenworte: Hefte für den Disput über Wissen, hrsg. von der Berlin-brandenburgischen Akademie der Wissenschaften. Eine stark gekürzte Fassung erschien am 11.8.2010 in der Frankfurter Allgemeinen Zeitung, S. N5 und am 15.8.2010 in FAZnet: http://www.faz.net/s/RubC3FFBF288EDC421F93E22EFA74003C4D/Doc~E45F43E9F45B9446FAEE9B814236CAB29~ATpl~Ecommon~Scontent.html.

2 Es wird im Text zugunsten der Lesbarkeit und aus sprachlichen Gründen nur die männliche Form gewählt.

in der Zeitschrift für Evaluation*, im* Evaluation Review *oder* American Journal of Evaluation. *Der Autor sollte größere wissenschaftliche Disziplin und Faktenbezogenheit zeigen."*

Ich habe verstanden und komme zu meinem Thema. Zuvor will ich aber noch sagen, dass ich den Mut der Deutschen Forschungsgemeinschaft (DFG) bewundere, dem Unwesen des "publish or perish" nunmehr entschieden entgegenzuwirken, indem Antragsteller in Zukunft nur noch maximal fünf Publikationen anführen dürfen. Jetzt wird endlich wieder weniger gezählt und mehr auf Inhalte geschaut. Ich wünschte mir, dass man auch bei den akademischen Evaluationen diesen Mut zur Umkehr hätte.

Nun ist ja das Gegenargument, dass die Verwendung öffentlicher Mittel überprüft werden müsse und es dafür Evaluationen brauche. Verwendung dürfe nicht Verschwendung werden. Das ist richtig, und daher bin ich auch nicht grundsätzlich gegen Überprüfungen. Ich bin nur gegen die Inflation der Überprüfungen. Selbst die Aufnahme eines Dreijährigen in den Kindergarten wird heute schon evaluiert. Und ich bin gegen das hohe Ausmaß, den Aufwand, die Bezahlung von Evaluatoren und die Überbürokratisierung der Verfahren, besonders durch die Akkreditierungsagenturen. Wenn ein Antrag für ein Exzellenzcluster 120 Seiten nicht überschreiten darf, aber die Akkreditierung eines BA-Studiengangs über 250 Seiten abverlangt, wenn heute fast mehr Zeit für Begutachtungen als für Forschung oder Lehre aufgewendet wird, dann stimmt etwas nicht mehr. Daher reihe ich mich in den Kreis derer ein, die da rufen: Hört auf mit dieser unverhältnismäßigen Torheit! Kehrt um! Besinnt Euch!

Wieder meldete sich hier der Evaluator: Der Autor argumentiert zunehmend hochschulpolitisch und polemisch. Es fragt sich, ob er überhaupt noch zum Thema kommen will.

Bevor ich es mir endgültig mit meinem Evaluator verderbe, will ich lieber Folge leisten, obgleich das, was ich zu sagen habe, mit meinen Bedenken gegen bestimmte und zu viele Evaluationen zu tun hat. Nicht selten nämlich, so scheint es, sind Evaluationen akademische Rituale, bei denen Ereignisse inszeniert werden, die dem Zweck des Ganzen, der Verbesserung der Forschung, allenfalls indirekt dienlich sind. Aber diese These kann ich erst überprüfen, nachdem ich kurz den begrifflichen Rahmen abgesteckt habe, in dem ich mich bewege. Dabei gilt es, die Begriffe „Evaluation" und „Ritual" festzulegen.

Unter „Evaluation" verstehe ich die nachträgliche Überprüfung einer Institution oder Person durch von außen kommende Experten. Unter „akademischem Ritual" verstehe ich solche Überprüfungen in einem universitären Bereich. Es fallen also keine Prüfungen wie Promotionen oder Habilitationen darunter oder Bewerbungsvorträge, die man alle auch als akademische Rituale auffassen kann, bei denen aber die Experten in der Regel nicht von außen kommen.

Ich konzentriere mich nachfolgend allein auf inszenierte Überprüfungen, im Besonderen die sogenannten Begehungen. Ich könnte auch die Akkreditierung nehmen, von der selbst der baden-württembergische Wissenschaftsminister Peter Frankenberg (CDU) sagt, dass sie „in ihrer bisherigen Form gescheitert sei", weil die Verfahren zu sehr an Formalien und bürokratischen Regulierungen und zu wenig am fachlichen Inhalt ausgerichtet seien (Schmoll 2009). Ich könnte auch die Habilitation nehmen, bei der es nicht allein um die wissenschaftliche Leistung eines Habilitanden geht, sondern bis zu einem gewissen Grad auch um die Frage, ob er in Auftreten und Erscheinung professorabel ist – was immer das heißen mag (vgl. etwa Breuer 1989).

Ich beschränke mich aber auf die Begehung. Der Begriff „Begehung" kommt aus dem Bauwesen oder von der Arbeitssicherheit, wo man eine Baustelle oder einen Arbeitsplatz überprüft. Der Begriff ist also treffend, denn tatsächlich geht es bei universitären Begehungen auch oft um die Sicherung von Arbeitsplätzen – jedenfalls aus der Perspektive des Mittelbaus.

Ich bin froh, dass sich an dieser Stelle der Evaluator nicht wegen erneuter unpassender Bemerkungen eingemischt hat, und kann daher fortfahren und bestimmen, was ich unter Ritualen verstehe

(vgl. z.B. Michaels 2003). Dies ist schon etwas komplexer. Ich konzentriere mich auf vier Merkmale und wende sie sogleich auf universitäre Begehungen an.

Das erste Merkmal betrifft die Verkörperung. Rituale setzen handelnde Personen voraus. Wer nur denkt oder fühlt, begeht kein Ritual. Das Ritual, wie ich es verstehe, setzt auch mehrere Personen voraus. Wer also nur ein Gutachten am Schreibtisch schreibt, begeht kein Ritual. Wohl aber bei Begehungen, also den Evaluationen vor Ort. Hier kommen Gruppen zusammen und verkörpern sich zu Gremien von Gutachtern und Begutachteten.

Das zweite Merkmal ist die Förmlichkeit. Rituale bestehen aus standardisierten, (mitunter stereotyp und redundant) wiederholten, somit nachahmbaren (und insofern öffentlichen) Handlungen. Sie sind aus Elementen (Ritemen) nach Regeln bewusst zusammengesetzt und fügen sich zu Ritualkomplexen (Sub- und Hauptritualen) zusammen. Dieses Regelwerk ist oft in Skripten oder Ritualhandbüchern festgehalten.

Dies alles trifft in hohem Maße auf universitäre Begehungen zu. Denn auch dabei sind die Handlungen vorgegeben. Es gibt ein genaues, teilweise vorgegebenes, teilweise stillschweigend von anderen Begehungen übernommenes Programm, dessen Struktur weitgehend festgelegt und als Skript auf der Homepage der DFG einsehbar ist.[3] So beginnt die Begehung eines Sonderforschungsbereichs am ersten Tag mit der Begrüßung durch den Senatsvertreter der DFG. Dann folgen ein Bericht des Sprechers, exemplarische Darstellungen von Forschungsprojekten, eine vom Senatssprecher geleitete Diskussion und ein Abschlusswort des Sprechers. Am Nachmittag schließen sich Besuche der Gutachter bei den einzelnen Teilprojekten an. Am Abend zieht sich die Gutachtergruppe zu ersten Beratungen zurück. Der nächste Tag beginnt mit der Plenarsitzung und einer Stellungnahme der Hochschule, meist des Rektors oder des Präsidenten. Die Anwesenheit des Kanzlers und eines Vertreters des Ministeriums ist vorgeschrieben. Der Plenarsitzung folgen Nachfragen der Gutachtergruppe und eine abschließende interne Beratung sowie die Beschlussfassung der Gutachter. Am Schluss wird der Sprecher dazu gebeten und das Urteil verkündet.

Dies ist der förmliche, vorgeschriebene Ablauf. Er entspricht rationalen, zweckgebundenen Vorgaben. Das wissenschaftliche Programm muss dargestellt, die Anträge auf Konsistenz und Genauigkeit geprüft, die finanziellen Forderungen nachgerechnet, der Anteil der Hochschule und des Landes besprochen werden. Man wüsste nicht, wie man es anders oder besser machen könnte.

Ebenso förmlich sind aber andere Richtlinien, die mehr oder weniger explizit gemacht werden. Es gibt eine gewisse Kleiderordnung, es gibt Namens- und Tischschilder. Die Bewirtung der Gutachter wird abgesprochen. Im erwähnten DFG-Vordruck heißt es: „Einen Mittagsimbiss am ersten Tag, ein Abendessen während der Klausur am ersten Tag und einen Mittagsimbiss am zweiten Tag. Bitte denken Sie auch an die Bereitstellung vegetarischer Speisen, Obst und ggf. Süßes sowie ausreichend Warm- und Kaltgetränke. Wir bitten darum, von der Bereitstellung alkoholischer Getränke Abstand zu nehmen." Die Anordnung der Stühle ist geregelt („Es ist wichtig, dass die Tisch- und Sitzanordnung eine Diskussion ermöglicht. Hörsäle eignen sich daher nicht gut für die Plenardiskussion"), ebenso die Begleitung der Gutachter („Es erleichtert die Anfahrt erfahrungsgemäß, wenn der Sonderforschungsbereich an beiden Tagen morgens für die Fahrt bzw. Begleitung vom Hotel zum Sitzungsort sorgt") oder das Hotel („Einzelzimmer mit Dusche/WC ... und gutem Qualitätsstandard, der eine erholsame Nachtruhe garantiert").

Diese Bestimmungen sind nur teilweise zweckrational, eher atmosphärisch. Gewiss, ein voller Bauch oder alkoholisierte Gutachter oder ein lautes, unwirtliches Hotel sind einer sachlichen Erörterung der anstehenden Fragen nicht gerade dienlich. Dies aber indirekt festzulegen bedeutet mehr. Es umgibt den Anlass mit einer gewissen Auratisierung.

Tatsächlich haben solche Begehungen ja ein wenig mit mittelalterlichen Herrschertreffen zu tun,

3 DFG-Vordruck 60.022 – 5/10.

denn schließlich begegnen sich Herrscher „von gleichem Stand“: Professorinnen und Professoren, Institutsleiter mit ihrem Gefolge, halt nur unter ritualisierten, das heißt förmlichen Bedingungen. Die Zeremonienmeister, die Mitarbeiterinnen und Mitarbeiter der DFG, halten sich dabei übrigens, wie es sich gehört, auffallend zurück: Man darf ihnen die Hoheit über die Organisation nicht anmerken, weil dies die Handlungsmacht der Evaluierten schmälern würde.

Ein entscheidender Aspekt der Förmlichkeit ist der formelle Beschluss zur Durchführung des Rituals (*intentio solemnis*), bei Begehungen durch förmliche Einladungen ausgedrückt. Er macht oft erst aus Alltagshandlungen Ritualhandlungen, indem er sie als besondere Handlungen kenntlich macht. Bloßes Wachbleiben wird so zu einer rituellen Nachtwache, bloßes Nichtessen zu religiös motiviertem Fasten. Ohne diesen förmlichen Beschluss stellt sich kein religiöses Verdienst ein. Eine spontane Evaluation wäre kein Ritual, auch wenn sie sich ritueller Elemente bediente. Rituale haben daher oft einen durch Zeichen (Glocken, Gesten, Kleidungswechsel etc.) signalisierten Beginn („Hiermit eröffne ich...“), mit denen die Abgrenzung zwischen Alltagswelt und Ritualwelt markiert wird. Diese Rahmung ist ein Teil der Förmlichkeit. Das Programm einer Begehung verdeutlicht dies eindrücklich.

Rituale bilden – dies ist das dritte Kriterium – einen bestimmten Modus des Handelns und unterscheiden sich darin, in welchem Maße sie sich auf ein Subjekt, auf die Gemeinschaft und auf eine überhöhte, transzendente Welt beziehen. Die Mischung dieser Eigenschaften macht den Modus eines Rituals aus. Bei einer Initiation wäre der auf das Subjekt bezogene Anteil höher, bei einer Begehung der auf die Forschergemeinschaft bezogene Anteil. In jedem Fall müssen aber alle drei Modi gegeben sein.

Der vielleicht umstrittenste Modus ist die Überhöhung oder Heiligung der Ritualhandlungen, mit denen sie zu einer anderen, meist als höher bewerteten Welt oder Tradition oder auf (heilige) Anfänge in Beziehung gesetzt werden. Dies geschieht etwa dadurch, dass bestimmte Handlungen als von Göttern vorgegeben angesehen werden.

Auch für eine Begehung trifft das Kriterium der Überhöhung zu – und zwar in einem entscheidenden Maße, denn es macht die Evaluation überhaupt erst zu einem akademischen Ritual. Worin liegt aber die Überhöhung? Bei Begehungen zeigt sie sich in dem Aufwand und in den verweisenden Bezügen. Eine Begehung ist ein großes, bedeutendes Ereignis, das inszeniert und zelebriert wird. Die Besten der Institution kommen zusammen und präsentieren sich im besten Licht. Es geht um Spitzenforschung, heute gern Exzellenz genannt. Wer Spitze sein will, muss sich gegenüber anderen überhöhen und absondern. Er darf eben nicht mehr nur Durchschnitt sein. Hierfür werden sachliche Gründe geltend gemacht, eben die Evaluationskriterien; auch diese sind in Katalogen der DFG festgehalten.

Freilich geht es um mehr. Es geht auch um Imponiergehabe und Macht. Beide Seiten müssen sich wechselseitig bestätigen, wie herausragend sie sind. Die Evaluatorenseite hat die Macht; das fordert Unterwerfungsgesten der Evaluierten heraus. Kritik an Gutachtern kommt nicht gut an, selbst wenn diese grobe Unkenntnis zeigen, selbst wenn bemerkt wird, dass sie weder die Forschungsberichte noch die anderen Unterlagen wirklich gelesen haben oder vom Thema wenig verstehen. Die Macht der Evaluatoren drückt sich nicht nur darin aus, dass sie über die Zusprechung oder Verweigerung von Mitteln befinden, sondern auch dadurch, dass sie Herren des Geschehens sind. Sie können eine Evaluation als freundlichen Fingerzeig auf Verbesserungen oder als „Hinrichtung“ von Kollegen gestalten. Sie haben die Agency über den Ablauf, die Zuteilung von Redezeit, die korrekte Durchführung der Evaluation, also über das rituelle Geschehen.

Mit diesen Maßnahmen, vor allem aber mit dem inszenierten Ereignis der Begehung wird die Überhöhung erreicht. Es geht um etwas Großes, und in diesen Tagen ist das Große: Viel Geld. Die Empfänger müssen sich würdig zeigen. In erster Linie, weil sie qualifiziert sind, aber auch indem sie sich hoheitsvoll verhalten. Das Ereignis muss kulturelle, symbolische Ordnungszeichen einsetzen, die ihre Überhöhung oder Solennität ausmachen. Als kulturelle Ordnungszeichen der Überhöhung können etwa Herrschaftszeichen (der Rektor erscheint mit Dienstwagen oder Plakette) oder

Insignien der Tradition (die Begehung findet in einem besonders ehrwürdigen Raum statt) gelten, mit denen zu Idealen (die Universität als höhere Bildungsinstitution, der Fortschritt der Forschung) oder Überpersönlichen Wert- und Ordnungsvorstellungen (die Universität als *universitas magistrorum et scholarium* oder *alma mater studiorum*) Bezug genommen wird. Akademische Evaluierungen sind kein TÜV. Sie verleihen sich Geltung und Autorität, weisen über die Gegenwart hinaus.

Kurz noch zum vierten Merkmal: Mit Ritualen geschieht etwas, das nicht trivial ist. Mit ihnen wird eine Transformation begangen. Nachher ist nicht gleich vorher. Das ist besonders bei Übergangsritualen augenscheinlich, weil bei einer Initiation ein Knabe zu einem heiratsfähigen Mann oder bei einer Hochzeit aus Mann und Frau ein Ehepaar wird. Aber auch in einer akademischen Evaluation wird der Status der Evaluierten verändert. Man wird zu einem Drittmittelempfänger und diese Aufwertung gehört fortan in den Lebenslauf wie Veröffentlichungen oder Preise.

So kann man also sagen, dass Evaluationen prozedurale akademische Rituale sind, weil sie den Kriterien der Verkörperung und Förmlichkeit folgen, im Modus einer selbstinszenierten Überhöhung begangen werden und statusbezogene Transformationen der Evaluierten bedeuten. Das alles macht nur begrenzt tauglich für kreatives Forschen. Rituale sind standardisierte Ereignisse mit einem Baukasten an vorgeschriebenen Handlungseinheiten, die Komplexität und Unwägbarkeit, die vielleicht wichtigsten Kriterien von Forschung, reduzieren helfen sollen. Sie sind nicht instrumental und können daher nicht leicht abgeändert werden. In diesem Sinn bedeutet die Ritualisierung von Evaluationen in der Tat Routine, Erstarrung und eine Loslösung von den ursprünglichen und eigentlichen Zielen der Forschung. Mit anderen Worten, das Ritual einer Begehung folgt nicht allein den Kriterien der objektiven wissenschaftlichen und zweckrationalen Beurteilung von Forschungs- oder Lehrleistungen. Akademische Evaluationen folgen der Vorstellung, dass man Wissenschaft in vergleichbare und damit messbare Einheiten zerlegen kann. Aber herausragende Ideen in der Forschung entstehen nicht in universitären Manufakturen, sondern in unvergleichlichen Köpfen, beim Spielen, Ausprobieren, Testen, oft genug als Zufallsprodukt, fast nie geplant. Kluge Evaluatoren wissen das; die DFG ermuntert zu einer solchen Klugheit.

Mein Evaluator bemerkt an dieser Stelle: Die Darstellung einer Begehung als akademisches Ritual überzeugt nur mäßig. Obgleich der Autor Ergebnisse der Ritualforschung gut anwendet, bleibt er doch den Beweis schuldig, dass eine so geeignete Überprüfung wissenschaftlicher Ergebnisse nicht doch zweckrational und somit das geeignete Mittel der Wahl ist. Die Frage stellt sich ja, wie groß der Ritualanteil am Gesamtgeschehen ist und ob er nicht vernachlässigenswert und unvermeidlich ist.

Da hat der Evaluator, der natürlich aus Gründen der Überhöhung anonym bleiben muss, obgleich er meiner Peergroup entstammt, Recht. Die Frage ist tatsächlich, ob es eine Alternative zu solchen Formen der Evaluation gibt. Und sind nicht die Vorteile einer ritualisierten Evaluation mit den Nachteilen abzuwägen?

Gewiss, Rituale bilden ein Vertrauenskapital (vergleichbar mit Pierre Bourdieus symbolischem, kulturellem und sozialem Kapital). Dieses sorgt für die Stabilität sozialer, politischer und wirtschaftlicher Beziehungen, es beseitigt Unsicherheit über die Legitimität der Ausgaben von öffentlichen Mitteln. Rituale vermitteln diese ersehnte Sicherheit, schaffen Vertrauen und verhindern oder vermindern Willkür, Beliebigkeit, Kontingenz, Komplexität und Individualität. Sie klammern die Sinn- bzw. Bedeutungsfrage weitgehend aus und werden zu einer Gewohnheit, bei der das richtige und angemessene Verhalten nicht jedes Mal neu ausgehandelt oder ins Bewusstsein gerufen werden muss. Dies kann für den Einzelnen eine Entlastungsfunktion bedeuten, aber auch für Institutionen wie DFG oder Universität eine effektive Ordnungsstruktur bilden.

Zudem haben Rituale oft genug die Autorität einer Tradition, einer Person, Organisation oder Institution hinter sich. Sie verbinden die Gegenwart mit der Vergangenheit, das Individuum mit der Gemeinschaft. Sie sind vielfach für eine Gruppe konstitutiv und fordern eine gebührliche Achtung bzw. Verweisung auf anerkannte Kontexte, Personen oder Institutionen.

Wer versucht, dies ändern zu wollen, indem er die Regeln von Ritualen vorsätzlich übertritt oder im Kern verändert, geht ein hohes Risiko ein, denn er kann – je nach Schwere des Regelbruchs – bestraft, geächtet oder ausgelacht werden. Das gilt auch für akademische Rituale. Man stelle sich einen SFB-Sprecher bei der Begehung im Trainingsanzug vor.

Ist also das Fazit, dass es unter Umständen keine Alternative zu den ritualisierten Evaluationen gibt? Ich fürchte ja, denn die Ausrichtung allein an sachgemäßen Kriterien würde die Überhöhung, mit der maßgeblich der Einsatz der Mittel begründet ist, nicht ermöglichen. Die Effizienz liegt in der Inszenierung von Misstrauen, aber Evaluierte und Evaluierende gehören der gleichen Kaste an und so sehr sie einander misstrauen, so sehr bestätigen sie sich gegenseitig. Längst orientiert sich die Wissensgenerierung nicht mehr allein an denen, die sie finanzieren: den Steuerzahlern.

Natürlich hat mein Evaluator auch diese Äußerung rot angestrichen. Aber dass er am Rande bemerkt, „der Autor scheint nicht wirklich zu wissen, worauf er hinaus will und bleibt im Unklaren, weil er Evaluierungen als Rituale sowohl verteidigt als auch als zu starre Instrumente kritisiert", verwundert mich doch sehr.

Ich muss diesen Punkt also verdeutlichen. Ohne Zweifel werden Rituale meist als starre und stereotype Handlungsmuster angesehen, doch wird dabei die Dynamik von Ritualen meist übersehen. Dies haben wir in unserem Heidelberger Sonderforschungsbereich „Ritualdynamik" immer wieder festgestellt. Rituale folgen einer Strukturdynamik, Geschichtsdynamik, Sozialdynamik und Erfahrungsdynamik. Das heißt, sie verändern ihre Strukturen, sie verändern sich in ihrer historischen Entwicklung, durch die wechselnde Zusammensetzung der sozialen Gruppen und durch Erfahrungen, die der Einzelne in ihnen macht. Entgegen verbreiteten Vorstellungen provozieren Rituale zum Beispiel Varianz. Trotz aller Förmlichkeit erzwingen situative und andere Faktoren eine stete Anpassung. Auch wenn diese Änderungen oft geleugnet werden, sind sie doch bei genauem Hinsehen bemerkbar.

So passt die DFG ihr „Ritualhandbuch" der Begehung immer wieder an. Sie reagiert damit auf unerwünschte Fehlentwicklungen. Aber es sind im Grunde die Rituale selbst, die das Neue provozieren. Im Grunde begleitet die Ritualkritik jedes Ritual schon mit seiner Entstehung. Denn da Rituale den Odem der sinnentleerten Starre vor sich her tragen, eignen sie sich besonders gut zur Kritik an überkommenen Traditionen. Da reicht schon eine prägnant artikulierte Kritik, um ganze Systeme einstürzen zu lassen, indem sie das Missverhältnis von Anspruch und Wirklichkeit auf den Punkt bringen. Paradigmatisch dafür der Spruch „Unter den Talaren – der Muff von tausend Jahren", mit dem 1967 die alten akademischen Rituale fast zum Erliegen kamen.

Eine Evaluation muss also keineswegs nur als schädlich angesehen werden. Sie kann gerade als Ritual Anlass geben, die Angemessenheit der Verfahren immer wieder zu überprüfen. Ich bin sicher, dass dies auch mit der neuen Welle der Evaluationen geschehen wird und insofern zuversichtlich. Vielleicht werden ja bald Studierende wieder ein Transparent vor den Professoren mit dem Spruch „Stellt Euch vor, es ist Akkreditierung und keiner geht hin" tragen.

Und was hätten wir für Alternativen? Ältere Methoden der Evaluation sind nicht mehr anwendbar. Die Inquisition, das erste professionalisierte Prüfungsverfahren, fällt aus. Die altindischen Redewettstreite, bei denen es – wie es in einer Upanischad heißt – soweit gehen konnte, dass man sein Gegenüber nicht überfragen durfte, weil sonst dessen Kopf zerplatzen würde, würden zwar die Bereitschaft zu Evaluationen deutlich verringen, nicht aber unbedingt zu einer Verbesserung der Forschungsleistungen führen.

Also bleibt mir nur der Appell zu mehr Vertrauen. Ich bin überzeugt, dass mehr Vertrauen nicht zu mehr Missbrauch führt. Evaluierungen ja, aber im Sinne der Überprüfung der Individualität von Forscherpersönlichkeiten, also mit einer sehr sorgfältigen Auswahl derer, denen Verantwortung gegeben wird. Meinetwegen auch Evaluationen der Ergebnisse von Zeit zu Zeit, in einem schön ritualisierten Rahmen, der das Potenzial des Ritualdesigns ausschöpft. Wettbewerb: Unbedingt,

Anreizsysteme auch, aber bitte nicht zu viele Evaluationen; das Syndrom der Evaluitis hat uns ja bereits wie die Schweinegrippe erfasst. Es ist selbstindiziert und da hilft nur, dass man dem Hype nicht zuviel Beachtung schenkt. Das Problem wächst sich dann von alleine wieder aus. Hoffentlich!

Zum Schluss danke ich meinem Evaluator – also mir selbst. Denn mein Denken ist eigentlich immer eine innere Begehung meiner selbst – und das kostet nicht mal etwas. Und ich danke dem Leser für seine prüfende und in diesem Sinne evaluierende Aufmerksamkeit.

Literatur

Breuer, Reinhard, 1989: Die physikalische Ständegesellschaft – Beschreibung einer Initiation, in: Kursbuch 97, September 1989, 137-149.

Floden, Robert E. / Weiner, Stephen S., 1978: Rationality to Ritual: The Multiple Roles of Evaluation in Governmental Processes. Policy Sciences 9 (1978), 9-18.

Frey, Bruno S., 2007: Evaluierungen, Evaluierungen ... Evaluitis. Perspektiven der Wirtschaftspolitik 8 (2007), 207-220.

Michaels, Axel, 2003: Zur Dynamik von Ritualkomplexen. Forum Ritualdynamik - Diskussionsbeiträge des SFB 619 der Ruprecht-Karls-Universität Heidelberg, Heft 3, 14.
Online: http://www.ub.uni-heidelberg.de/archiv/4583. Nachdruck in: „Ritualbegriff und Ritualanalyse. Beiträge des Workshops vom 30./31. Oktober 2003 in Konstanz", hrsg. vom SFB 485 „Norm und Symbol", Arbeitspapiere Nr. 47, Juli 2004.

Power, Michael, 1997: The Audit Society. Ritual of Verification. Oxford: University Press.

Schmoll, Heike, 2009: Ärgernis Akkreditierung. Frankfurter Allgemeine Zeitung, Nr. 291, vom 15. Dezember 2009, 10.

Schwarz, Christine, 2006: Evaluation als modernes Ritual. Zur Ambivalenz gesellschaftlicher Rationalisierung am Beispiel virtueller Universitätsprojekte. Hamburg: Lit Verlag.

Meike Olbrecht, Thamar Klein

SFB-Begutachtung: Entscheidungsfindung in Gruppen

1 Einleitung

Ziel des vorliegenden Papiers ist es, einen Einblick in das Projekt „Peer Review in der DFG: Panelbegutachtung am Beispiel der Sonderforschungsbereiche“[1] zu bieten, welches die Begutachtung durch Gutachter/innengruppen untersucht. Dabei werden neben Verlauf und ersten vorläufigen Ergebnissen des Projekts die genutzten Methoden vorgestellt.

Generell lassen sich bei der Begutachtung eines Forschungsantrags durch mehrere Fachkolleginnen und Fachkollegen zwei Vorgehensweisen unterscheiden: Bei der schriftlichen *Einzelbegutachtung* beurteilen die Wissenschaftlerinnen und Wissenschaftler *unabhängig* voneinander einen Antrag und senden ihre Beurteilungen in Form von Gutachten an die Organisation, die die Förderentscheidung trifft. Beim *Panel Peer Review* wird vor der Entscheidung ein Antrag in einer Gutachtendengruppe *gemeinsam* beraten und beurteilt. Während bei der schriftlichen Einzelbegutachtung der oder die Entscheidungsträger/in einer Forschungsförderungsorganisation die Urteile und Kommentare mehrerer Gutachtenden zu einer Entscheidung zusammenführt, liegt beim Panel Peer Review ein gemeinsames Urteil von allen Gutachtenden als Entscheidungsvorlage vor. Das Panel Peer Review unterscheidet sich demnach von der schriftlichen Einzelbegutachtung vor allem dadurch, dass die Gutachtenden im Panel Peer Review *durch einen diskursiven Prozess* zu einem gemeinsamen Urteil kommen (Olbrecht/Bornmann 2010).

Weltweit werden immer größere Summen der Forschungsförderung über Beurteilungen durch Gutachtendengruppen anstatt über Einzelbegutachtung vergeben. Entgegen dieser Entwicklung konzentriert sich die Peer Review-Forschung bisher hauptsächlich auf die Einzelbegutachtung und fragt nach der Reliabilität, Validität und Fairness dieser Verfahren. Der Begutachtung in Panelsitzungen wurde bisher wenig Aufmerksamkeit geschenkt (Olbrecht/Bornmann 2010).

Die Peer Review-Forschung der letzten Jahre hat sich zum größten Teil mit dem Peer Review von Zeitschriften beschäftigt (vgl. Überblick in DeVries/Marschall/Stein 2009; Overbeke/Wager 2003/ Weller 2001); seltener wurde das Peer Review von Forschungsanträgen untersucht (Daniel/Mittag/ Bornmann 2007; Lamont 2009; Wessely 1998). Das Peer Review bei Forschungsanträgen widmete sich vor allem der Fairness und prognostischen Validität der Urteile einzelner Gutachtender bzw. der finalen Entscheidung einer Förderorganisation. Wir konnten lediglich einige wenige Studien recherchieren, die sich mit den Prozessen der Urteilsfindung in einem Panel beschäftigt haben (u.a. Johnson 2008; Langfeldt 2001, 2004; Obrecht/Tibelius/D’Aloisio 2007). Andere Forschungsgebiete, vor allem die sozialpsychologische Forschung, haben sich dagegen bereits intensiv damit beschäftigt, wie eine Gruppe von Personen zu einem gemeinsamen Urteil kommt und welche Phänomene dabei auftreten können, die einen (unerwünschten) Einfluss auf das Urteil haben (Olbrecht/Bornmann 2010).

Die geringe Zahl empirischer Studien zur Panelbegutachtung im Bereich der Peer Review-Forschung ist angesichts der methodischen Schwierigkeiten und der problematischen Feldzugänge zu Begutachtungssitzungen erklärlich. Gleichwohl ist es angesichts der zunehmenden Bedeutung dieses Begutachtungstyps erstaunlich, dass über Mechanismen der Konsensfindung und der Leistungsfähigkeit sowie über potentielle Bias-Faktoren dieser Form kollektiver Beurteilung so wenig bekannt ist.

1 Das Projekt “Peer Review in der DFG: Panelbegutachtung am Beispiel der Sonderforschungsbereiche” wird am Institut für Forschungsinformation und Qualitätssicherung durchgeführt. FÜr weitere Informationen siehe unter http://www.forschungsinformation.de

1.1 Ziele und Fragestellung des Projekts

Das Projekt „Peer Review in der DFG: Panelbegutachtung am Beispiel der Sonderforschungsbereiche" fragt danach, welche Effekte bei Panelbegutachtungen auftreten können. Ziel des Projekts ist es, den gruppendynamischen Prozess der Urteilsfindung von Gutachtendengruppen zu untersuchen. Zu diesem Zweck sollen unter anderem folgende Fragestellungen bearbeitet werden: Wie finden Gutachtendengruppen zu einem gemeinsamen Urteil? Welche gruppendynamischen Prozesse lassen sich beobachten? Worin bestehen diese? Welchen Einfluss hat die Gruppendiskussion auf das individuelle Urteil? Wie bewerten die an der Begutachtung beteiligten Akteure und Akteurinnen den gemeinsamen Urteilsfindungsprozess?

Das vorliegende Projekt untersucht diese Fragen anhand der Gruppenbegutachtung von Sonderforschungsbereichen (SFB). Das Projekt gliedert sich in eine explorative Phase und eine Hauptphase. Die hier vorgestellten Ergebnisse und Methoden stammen aus der explorativen Phase und können deshalb auch nur vorläufigen Charakter haben. In der Hauptphase wurden die Erhebungsinstrumente noch einmal überarbeitet und es fand eine Konzentration auf die Wissenschaftsgebiete „Geistes- und Sozialwissenschaften" sowie „Lebenswissenschaften" statt.

2 Methodisches Vorgehen

Um unterschiedliche Ebenen und Aspekte des Untersuchungsgegenstandes zu erfassen, wurden im Projekt sowohl eine Reihe qualitativer (Expert/inn/en-Interviews, Leitfadeninterviews, nichtteilnehmende Beobachtung) und quantitativer Methoden (standardisierte Benotungsabfrage, Onlinebefragung[2], kognitive Verfahren wie Freelisting, Ranking, Pile-Sort) kombiniert als auch sämtliche Akteurinnen- und Akteurs-Perspektiven und alle Stufen des Begutachtungsprozesses verfolgt. Diese beinhalten das Beratungsgespräch, Erstbegutachtung, Folgebegutachtungen sowie die Sitzungen des Senats- und Bewilligungsausschusses.

Die qualitativen und kognitiven Methoden dienen dabei vor allem dazu, differenzierte gruppenspezifische Informationen über Beratungs- und Begutachtungssituationen zu gewinnen und Stärken sowie Schwächen des Systems zu erheben. Dafür werden die Erfahrungen und Fachkompetenzen aller beteiligten Akteurinnen und Akteure genutzt. Neben den Gutachtenden sind am Beratungs-, Begutachtungs- und Bewilligungsprozess die Mitarbeiter und Mitarbeiterinnen der DFG, die Antragstellenden, die DFG-Berichterstatterinnen und Berichterstatter[3] sowie Vertreterinnen und Vertreter der Hochschulleitungen und des Landes beteiligt. Die Einnahme dieser unterschiedlichen Perspektiven erlaubt die Kombination und Zusammenführung des Expert/inn/enwissens sehr unterschiedlicher Personengruppen. Die von uns durchgeführte sequentielle Verbindung unterschiedlicher Methoden, Techniken und Perspektiven bietet einen beträchtlichen Zuwachs an Erkenntnis und verringert die Gefahr, den „blinden Flecken" der jeweiligen Methoden bzw. der Akteure und Akteurinnen aufzusitzen. Es geht dabei darum Divergenzen aufzuzeigen und durch die Komplementarität der Resultate zu einem „kaleidoskopartigen" (Köckeis-Stangl 1982: 363) Modell von interagierenden Wirklichkeiten zu gelangen. Dies erscheint als besonders bedeutend im Zusammenhang mit möglichen Evaluationsunterschieden zwischen verschiedenen Wissenschaftskulturen sowie der Vielzahl der beteiligten Akteurinnen und Akteure.

Im Folgenden werden die Erhebungsmethoden des Projekts im Einzelnen vorgestellt.

2 Diese befindet sich noch in der Planungsphase.

3 Bei den beiden Berichterstattenden handelt es sich um Mitglieder des Bewilligungsausschusses für die Sonderforschungsbereiche. Sie müssen das Gesamtergebnis der SFB-Begutachtung dem Ausschuss vorstellen, der in letzter nstanz über die Förderung entscheidet.

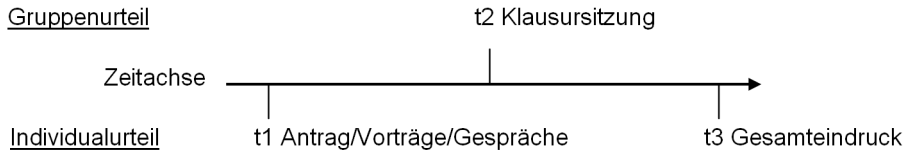
2.1 Standardisierte Benotungsabfrage

Bei der standardisierten Benotungsabfrage handelt es sich um eine Methode, um den Prozess der Urteilsbildung vom individuellen Urteil hin zum Gruppenurteil zu untersuchen.

Alle Gutachtenden wurden zu zwei verschiedenen Zeitpunkten gebeten, ein Individualurteil für den gesamten SFB und dessen einzelne Teilprojekte abzugeben (siehe Abbildung.1: *t1*, *t3*). Die Bewertung orientiert sich an der bei SFB-Begutachtungen üblichen Skala. Diese reicht von der Note „exzellent" bis zu der Note „nicht förderungwürdig". Dazwischen bestehen folgende Abstufungen: „sehr gut bis exzellent", „sehr gut", „gut bis sehr gut" und „gut". Für den Fall von Teilprojekten, bei denen sich die Gutachtenden noch keine abschließende Meinung gebildet hatten, bestand die Möglichkeit „keine Bewertung" anzukreuzen.

Zwischen den beiden Zeitpunkten *t1* und *t3* fällten die Gutachtenden in der Klausursitzung ein gemeinschaftliches Gruppenurteil als Teil des DFG-üblichen Begutachtungsprozederes (*t2*).

Abbildung 1: Zeitachse standardisierte Benotungsabfrage



Das Individualurteil zu *t1* wurde auf Grundlage der Anträge sowie der am Vormittag des ersten Begutachtungstages gehaltenen Kurzvorträge der Antragstellenden und den am Nachmittag folgenden Einzelgesprächen zwischen den Mitgliedern der Prüfungsgruppe und den einzelnen Arbeitsgruppen gefällt. Die Bögen von *t3* spiegeln die individuellen Urteile nach Abschluss der zwei Begutachtungstage wieder.

Ziel war es, den Prozess der Urteilsbildung vom individuellen Urteil auf Basis des Antrags, der Kurzvorträge und der Einzelgespräche über das Gruppenurteil in der SFB-Begutachtung bis zum individuellen Urteil nach der Gruppenevaluation zu untersuchen. Insbesondere interessierte uns, inwieweit der Mittelwert der Individualurteile zum Zeitpunkt *t1* dem konsensual gefällten Gruppenurteil ent- oder widersprach. Erste Ergebnisse hierzu finden sich in den Abschnitten 3.1 und 3.2.

Sozialpsychologische Experimente verweisen darauf, dass in Gruppendiskussionen von allen geteilte Informationen öfter wiederholt werden als nicht geteilte Informationen (vgl. Larson/Foster-Fishman /Keys 1994; Dennis 1996), dass Gruppenurteile zur Polarisierung neigen (die Gruppe nimmt eine extremere Haltung ein als der Durchschnitt der Individualurteile – für einen Überblick siehe auch Isenberg 1986) und dass Gruppen ihre Urteile oft auf der Basis der vorherigen individuellen Urteile und nicht auf der Basis ihres Gesamt-Know-hows fällen (Schulz-Hardt et al. 2006).

Bei der Untersuchung der zentralen Fragestellung des Projekts nach den bei Panelbegutachtungen auftretenden Effekten und gruppendynamischen Prozessen stellte sich die Frage, ob und inwieweit sich diese sozialpsychologischen experimentellen Beobachtungen in der nichtexperimentellen Situation des Beratungsgesprächs und der Klausursitzung nachweisen lassen würden. Der Abgleich der standardisierten Benotungsabfrage mit den Aufzeichnungen aus den nicht-teilnehmenden Beobachtungen der Sitzungen eröffnet die Möglichkeit, diesbezügliche Fragen zu beantworten, etwa von welchen Personen(gruppen) bestimmte Informationen eingebracht werden, wie lange die Gruppe über bestimmte Punkte diskutiert, welche Punkte aufgegriffen werden oder ob und wie oft abweichende Meinungen von Individuen in die Gruppe hineingetragen werden.

Ebenfalls war von Interesse, inwieweit das individuelle Urteil zu *t3* von dem zuvor gemeinsam

beschlossenen Urteil zu *t2* abwich. Starke Abweichungen könnten hier einen Hinweis auf unerwünschte Gruppenphänomene, wie beispielsweise Konformitätsdruck, geben.

2.2 Nicht-teilnehmende Beobachtung

Bei der in dieser Studie angestrebten Beobachtung handelt es sich um eine offene (die Beobachteten wissen von der Beobachtung), nicht-teilnehmende (die Beobachterinnen sind nicht Teil des Feldgeschehens), natürliche (Beobachtung im natürlichen Feld) Fremdbeobachtung (die Beobachtenden sind selber nicht Gegenstand der Beobachtung). Die Beobachtung ermöglicht ein Verstehen von Abläufen sozialer Prozesse, subjektiver Sichtweisen sowie (sub-)kultureller und sozialer Regeln. Wirklichkeit wird in diesem Zusammenhang als von individuellen oder kollektiven Akteur/inn/en hergestellt betrachtet. Die Interaktionen, durch die diese Wirklichkeit hergestellt wird und durch die Sinnstrukturen und Regeln vermittelt werden, sollen von innen heraus beobachtet und analysiert werden. In diesem Projekt ermöglicht diese Methode beispielsweise die Beobachtung möglicher Unterschiede in der Diskussions- und Evaluationskultur zwischen Geistes-/Sozialwissenschaften und Lebenswissenschaften. Weitere Vorteile der nicht-teilnehmenden Beobachtung bestehen unter anderem darin, dass durch die nicht-teilnehmende Beobachtung unbewusste Tatsachen – bzw. Fakten, die als von Insidern selbstverständlich und nicht erwähnenswert betrachtet werden sowie Gegebenheiten, die in Interviews aus anderen Gründen ungenannten bleiben – erfahren und gesammelt werden können.

2.3 Dokumentenanalysen

Wichtigste zu analysierende Dokumente in dieser Studie sind die gemeinsam von den DFG-Mitarbeitenden und Berichterstatter/inne/n verfassten Protokolle. Die Entscheidung über die Förderung von Sonderforschungsbereichen erfolgt im Bewilligungsausschuss auf der Grundlage dieser Ergebnisprotokolle, der Vor-Ort-Begutachtungen sowie der Kurzberichte der beiden bei der Begutachtung anwesenden Berichterstatter/innen im Bewilligungsausschuss.

Bei der Dokumentenanalyse steht die Auswertung der Aspekte des Begutachtungsverfahrens im Vordergrund, die als protokollierungswürdig erachtet wurden und die im letzten Schritt des Begutachtungsprozesses – dem Bewilligungsausschuss – relevant werden. Es interessieren hierbei der Vergleich und die Analyse der während der Klausursitzung diskutierten Kriterien mit den in das Protokoll eingegangenen Kriterien sowie den im Bewilligungsausschuss hervorgehobenen Kriterien.

2.4 Leitfadeninterviews und kognitive Methoden

Die Leitfadeninterviews gliedern sich in vier grobe Teilbereiche, in denen allgemeines Wissen und Erfahrungen mit Peer Review, allgemeines Wissen und Erfahrungen mit SFB-Beratungen/Begutachtungen und Fragen zur konkreten Beratung/Begutachtung behandelt sowie Kognitive Methoden eingesetzt werden. Dabei geht es um die Aufdeckung und Analyse differenzierter gruppenspezifischer Denkmodelle und Begriffssysteme, mit denen die unterschiedlichen Akteure und Akteurinnen beim Begutachtungsprozess operieren (kulturelle Domänen).

Eine kulturelle Domäne ist ein abgegrenzter Wissensbereich mit innerer Struktur, der von Angehörigen einer bestimmten „Kultur" geteilt wird (Borgatti 1994, Schnegg/Lang 2008). Dabei kann es sich bei der Kultur beispielsweise auch um Fachkulturen handeln. Kulturelle Domänen können zum Beispiel Begriffssysteme sein, mit denen Farben, Verwandtschaftstermini oder – wie im Falle dieser Studie – Kriterien für Begutachtungen sowie bei Begutachtungen auftretende Probleme klassifiziert werden. Bei der Gewinnung und Untersuchung dieses lokalen Wissens folgt unsere Studie den Erfahrungen aus der *Cultural Domain Analysis* (Alexiades/Sheldon 1998; Borgatti 1994; Bernard 2002) und setzt freelistings, pile sorts, rankings und true/false statements als quantitative Erhebungswerkzeuge ein,

die mittels *SPSS, Anthropac* (Borgatti 1996) und *VAP_Pilesorts* ausgewertet werden.

3 Darstellung erster vorläufiger Ergebnisse

Das Projekt gliedert sich in einen explorativen Teil und eine Hauptstudie. Ziel des explorativen Teils ist die Erarbeitung und Validierung von Erhebungsinstrumenten auf der Basis erster qualitativer Erhebungen und Ergebnisse, die im Anschluss, im Hauptteil des Projekts, zur Anwendung kommen. Die Ergebnisse des ersten Teils werden in der Hauptstudie spezifiziert und auf ihre Generalisierbarkeit geprüft. Ende Mai 2010 wurde die explorative Projektphase abgeschlossen. Bis zu diesem Zeitpunkt fanden drei Beobachtungen von Beratungsgesprächen und drei von SFB-Einrichtungsbegutachtungen statt. Darüber hinaus wurden rund 80 Leitfadeninterviews mit Gutachtenden, SFB-Berichterstattenden, Antragstellenden und DFG-Mitarbeitenden durchgeführt. Die Dauer der Interviews betrug im Durchschnitt eine Stunde.

In einer Art Werkstattbericht sollen im Folgenden einige erste Ergebnisse vorgestellt werden. An dieser Stelle muss explizit darauf hingewiesen werden, dass es sich um vorläufige Ergebnisse handelt. Inwiefern sich diese in der Hauptphase bestätigen, ist zum jetzigen Zeitpunkt noch offen.

3.1 Gruppenpolarisierung

In der Sozialpsychologie wird mit dem Begriff der Gruppenpolarisierung die Tendenz bezeichnet, im Anschluss an eine Gruppendiskussion eine extremere Position einzunehmen als vor der Diskussion, und zwar in die Richtung, in die der Durchschnitt der Einzelpositionen vor der Diskussion tendierte (Isenberg 1986; Moscovici/Zavalloni 1969). Für den Effekt der Gruppenpolarisation gibt es zwei wichtige Erklärungsansätze: (1) Theorie persuasiver Argumente (Burnstein/Vinokur 1977) und (2) der Erklärungsansatz des sozialen Vergleichs (Sanders/Baron 1977).

Zu (1): Der Erklärungsansatz der persuasiven Argumentation geht davon aus, dass sich eine individuelle Position aus der Anzahl, der Ausrichtung und der Überzeugungskraft der Argumente ergibt, über die eine Person verfügt. In einer Diskussion werden Argumente der verschiedenen Gruppenmitglieder ausgetauscht. Wenn die Mitglieder neue und glaubwürdige Argumente vortragen, die eine bereits bei einer Person dominierende Position unterstützen, kann dies dazu führen, dass der/die Einzelne aufgrund der zunehmenden Anzahl unterstützender Argumente bereit ist, eine extremere Position einzunehmen als vor der Diskussion (Burnstein/Vinokur 1977).

Zu (2): Der Erklärungsansatz des sozialen Vergleichs geht davon aus, dass Gruppenmitglieder dazu neigen, sich mit anderen Mitgliedern zu vergleichen. Sie haben das Bedürfnis, sich selbst positiv zu sehen und Zustimmung von anderen zu erlangen (Goethals/Zanna 1979). Gleichzeitig besteht der Wunsch, sich von den anderen zu unterscheiden. Wenn die Gruppenmitglieder feststellen, dass die anderen in der Tendenz dieselbe Meinung vertreten, sind sie bereit, sich von der Mehrheitsmeinung abzuheben, indem sie eine extremere Position einnehmen als die Mehrheit der Gruppe – in die Richtung, in die die Mehrheit der Gruppenmitglieder ohnehin tendiert (Myers 1978). Dadurch können sie erwarten, dass sie trotz ihrer abweichenden Meinung positiv von der Gruppe beurteilt werden.

Es gibt unterschiedliche Bereiche, in denen das Phänomen der Gruppenpolarisation zu beobachten ist. Nijstad schreibt dazu: "Group polarization has since been shown in a variety of contexts, including (…) jury decisions" (Nijstad 2009: 37). Wir können deshalb annehmen, dass sie auch im Panel Peer Review auftritt. Hier würde das Phänomen dazu führen, dass das Gruppenurteil im Anschluss an die Panelsitzung positiver oder negativer ausfällt als der Durchschnitt der Individualmeinungen vor der Diskussion (Olbrecht/Bornmann 2010).

Diese Vermutung wurde durch unsere Ergebnisse aus der explorativen Phase bestätigt. Mit Hilfe der Benotungsabfrage der SFB-Teilprojekte, welche die Gutachtenden vor und nach der Gruppendiskussion vornahmen, war es uns möglich, die Bewertung der Projekte zu drei verschiedenen Messzeitpunk-

ten miteinander zu vergleichen (vgl. 2.1). Bei der standardisierten Benotungsabfrage handelt es sich um eine Methode, um den Prozess der Urteilsbildung vom individuellen Urteil hin zum Gruppenurteil zu untersuchen. Der Vergleich dieser drei Urteile zeigt, dass die Gruppendiskussion zu einer stärkeren Polarisierung führt, wobei sich das Urteil in den meisten Fällen in eine positive Richtung verstärkt. Das heißt, sowohl das Gruppenurteil als auch der Durchschnitt der Einzelmeinungen nach der Diskussion fielen positiver aus als der Durchschnitt der Einzelmeinungen vor der Diskussion.

In den Interviews wurde, wenn es um die Benotung von Teilprojekten ging, teilweise ein strategisches Kalkül genannt, welches neben dem sozialpsychologischen Phänomen der Gruppenpolarisierung als Erklärung für den Trend zum positiveren Urteil angeführt werden kann: Die Gutachtenden gaben in den Interviews an, dass sie – sofern sie einen SFB für förderungswürdig hielten – positivere Noten für die einzelnen Teilprojekte vergaben, als sie es eigentlich für angemessen hielten. Dadurch wollten sie sicher gehen, dass der gesamte SFB auch letztlich gefördert wird. Im Folgenden sollen zwei Zitate dieses Phänomen des strategischen Aufwertens beispielhaft deutlich machen:

> *"Das ist schon eine Schwäche des DFG-Prozesses insgesamt. Durch diese sehr begrenzten Mittel und die Überbuchung hat ja kein Antrag eine Chance, wenn er nicht super herausragend und ohne den kleinsten Makel aus einer Begutachtung herausgeht. Und das macht es natürlich dann auch in der Endrunde ein bisschen schwieriger, wenn man sagt, eigentlich möchte ich, dass das Ding durchkommt. Weil ich halte das zusammenfassend für eine tolle Sache, tolle Leute, die da Antragsteller sind und so weiter. An der kleinen Ecke, da könnte man es noch so ein bisschen anders machen. Das traut man sich dann schon gar nicht mehr so richtig offen zu sagen." (Gutachter/in A)*

Ein/e weitere/r Gutachter/in gab an:

> *"Wenn die DFG im Voraus sagt, wir kriegen nur exzellente SFBs durch. Dann shiftet man das natürlich zur Exzellenz. Das ist dann vielleicht …im Einzelfall (…) mag das zu einer Inflation der guten Noten führen. (…)" (Gutachter/in B)*

3.2 Varianz der Individualurteile

Die Analyse der standardisierten Benotungsabfrage ergab Abweichungen zwischen den Einzelmeinungen zu den Zeitpunkten vor der Gruppendiskussion (*t1*) und nach dem Gruppenurteil (*t2*). Abbildung 2 zeigt die Bewertung eines Teilprojekts im Rahmen einer SFB-Begutachtung, welches von der Gutachtendengruppe nicht zur Förderung empfohlen wurde. Im Weiteren soll dieses einzelne Teilprojekt beispielhaft betrachtet werden.

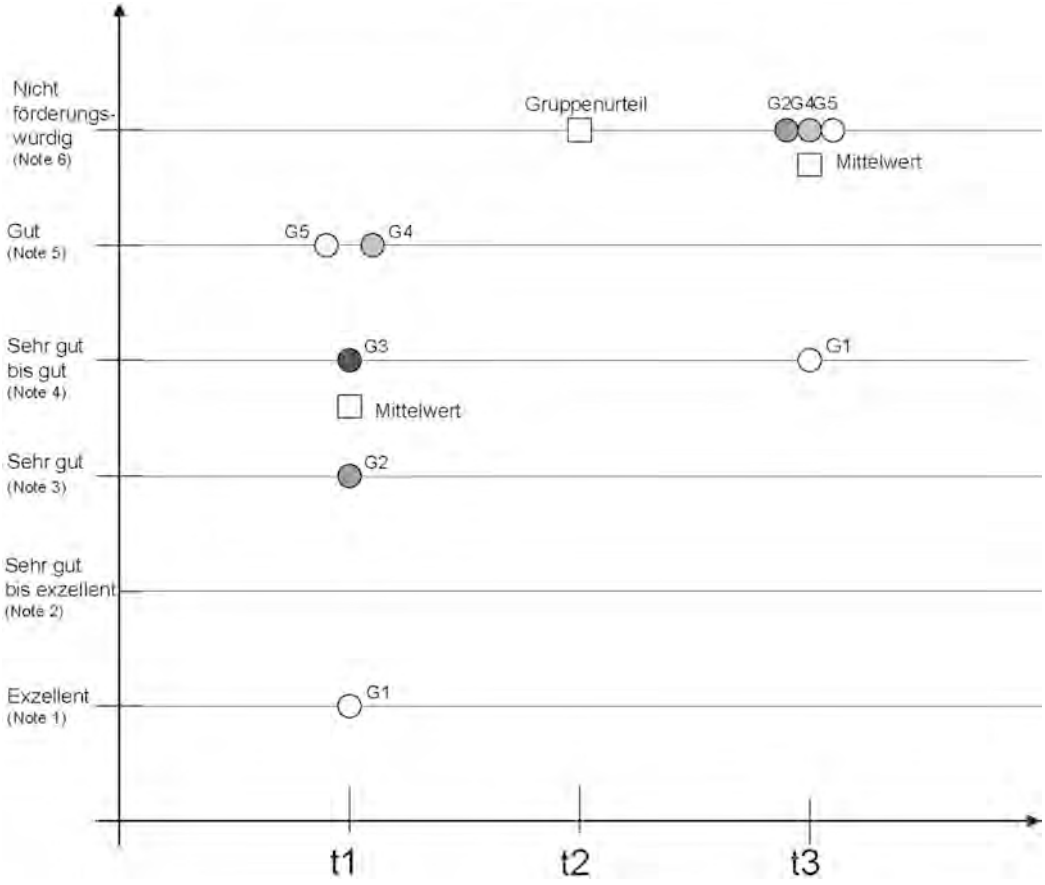
In Abbildung 2 sind jeweils die Individualurteile der einzelnen Gutachtenden vor *(t1)* und nach der Diskussion *(t3)* mit Punkten abgetragen. Der Mittelwert ist mit einem Quadrat symbolisiert. In der Mitte ist das konsensuale Gruppenurteil *(t2)* der Klausursitzung ebenfalls als Quadrat abgetragen. Bei dem dargestellten Teilprojekt bestand eine große Varianz bei den Einzelurteilen vor der Diskussion: Der/die Gutachtende G1 hat das Teilprojekt mit „exzellent" (Note 1) bewertet. Im Gegensatz dazu waren zwei andere Gutachtende (G4 und G5) der Meinung, es wäre lediglich ein „gutes" Projekt (Note 5). Der Mittelwert der Bewertung dieses Teilprojekts vor der Diskussion befindet sich bei der Note 3,6.

Nach der Diskussion zeigt sich ein einheitlicheres Bewertungsbild, das eine geringe Varianz in den Einzelurteilen aufweist. Lediglich der/die Gutachtende, der/die das Projekt vor der Diskussion mit „exzellent" benotete, ist auch nach der Diskussion der Meinung, dass es sich – wenn auch schlechter bewertet – um ein förderungswürdiges Teilprojekt handelt. Der/die Gutachtende G3 kreuzte nach der Diskussion „keine Bewertung" an, weswegen er/sie bei den Einzelurteilen nach der Diskussion nicht mehr abgetragen ist.

Aufgrund des Urteilsfindungsprozesses der Gutachtendengruppe ist das Teilprojekt im Er-

gebnis nicht zur Förderung empfohlen worden. Diese Entscheidung ist auf Grundlage des Durchschnittswerts der Einzelurteile nicht vorherzusehen gewesen. Es stellt sich die Frage, wie es zu diesem Urteil gekommen ist. Um diese beantworten zu können, ist es notwendig, den Diskussionsprozess genauer zu betrachten. Hier erweist sich die Methoden-Triangulation von Vorteil, da der Zugewinn an Wissen aus der nicht-teilnehmenden Beobachtung das Wissen aus der standardisierten Benotungsabfrage ergänzen kann. Aus dem Beobachtungsprotokoll geht hervor, dass es zwar nur wenig Diskussion um dieses Teilprojekt gab (insgesamt beteiligten sich nur drei Gutachtende daran), dass sich jedoch die beiden Vorstellenden darüber einig waren, dass es sich um ein schwaches Projekt handelt. Auf die Rolle der Vorstellenden innerhalb der Gruppe soll deshalb im Folgenden eingegangen werden.

Abbildung 2: Individualurteile versus Gruppenmeinung (abgelehntes Teilprojekt)



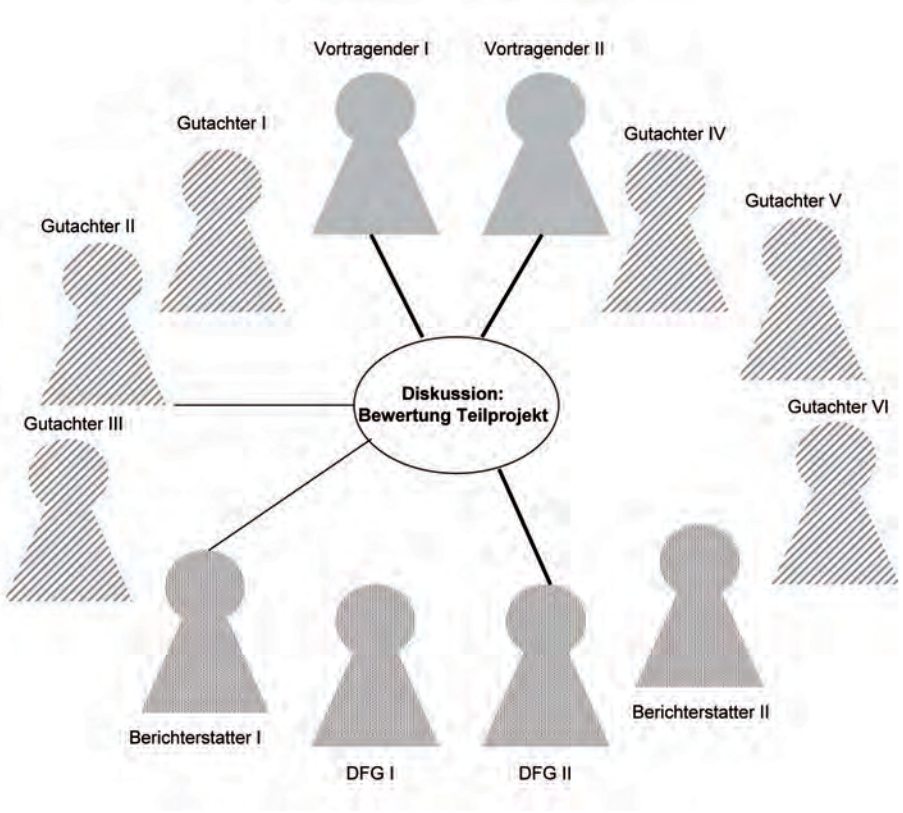
3.3 Gruppendiskussion

In Abbildung 3 ist die Diskussion des abgelehnten Teilprojekts dargestellt. Die Abbildung zeigt die Gutachtendengruppe bestehend aus acht Gutachtenden, zwei DFG-Mitarbeitenden und zwei Berichterstattenden.

Bei jedem Teilprojekt gibt es mindestens zwei Gutachtende, die sich besonders intensiv auf dieses vorbereitet haben. Sie werden vor der Sitzung von der DFG aufgrund ihrer besonderen Expertise in dem Fachgebiet des Teilprojekts ausgewählt. Im Weiteren werden diese Gutachtenden als *Vorstellende* bezeichnet. Diese Vorstellenden erläutern der Gutachtendengruppe ihre Bewertung des Projekts. In der Regel beginnt eine/r der beiden, indem er/sie kurz den Inhalt des Projekts wiedergibt, im Anschluss folgt seine/ihre Bewertung anhand verschiedener Kriterien, wie zum Beispiel Qualität des Projekts, Qualifikation des/der Teilprojektleiters/Teilprojektleiterin, zu

erwartender Erkenntnisgewinn, Einbindung des Projekts in den SFB-Verbund usw. In den meisten Fällen beendet der/die Vortragende seine/ihre Ausführungen damit, dass er/sie das Projekt auf der für den SFB üblichen Notenskala von „exzellent" bis „nicht förderungswürdig" bewertet. Der /die zweite Vorstellende schließt seine/ihre Ausführungen denen des Vorredners an. In Abbildung 3 sind die Vorstellenden des Projekts grau dargestellt. Die Linien verdeutlichen, wer sich mit Redebeiträgen an der Diskussion beteiligt hat. Je dicker die Linien, umso mehr Wortbeiträge wurden geäußert. Die Abbildung zeigt, dass sich außer den zwei Vorstellenden lediglich ein/e weitere/r Gutachtende/r sowie ein/e Berichterstatter/in mit wenigen Beiträgen an der gesamten Diskussion beteiligten. Während der Dauer von 20 Minuten, die die Diskussion über das Teilprojekt andauerte, beteiligten sich im Wesentlichen drei Personen an dem Urteilsfindungsprozess: die beiden Vorstellenden und ein DFG-Mitarbeitender. Zum Schluss kamen die beiden Vorstellenden zu dem Ergebnis, dass das Projekt nicht zur Förderung empfohlen wird. Die weiteren fünf Gutachtenden beteiligten sich demnach mit keinem Wortbeitrag an der Diskussion und Bewertung des Projekts.

Abbildung 3: Ablauf der Gruppendiskussion



Die Gruppendiskussion ist demzufolge zu einem anderen Ergebnis gekommen, als dies der Mittelwert der Einzelurteile nahe gelegt hätte, weil ein Großteil der Begutachtungsgruppe ihre Benotung des Projekts nicht äußerte. Da die Befragung der Gutachtenden hinsichtlich ihres Urteils vor und nach der Gruppendiskussion in der explorativen Phase anonym war, können wir nur darüber mutmaßen, dass die beiden Vortragenden aus Abbildung 3 jene Gutachtenden waren, die das Projekt in Abbildung 2 mit der Note 5 bewerteten. Das Beobachtungsprotokoll unterstützt diese Vermutung, denn es zeigt, dass die beiden Vorstellenden sich schnell einig darüber waren, dass es sich um ein schwaches Projekt handele. Die Diskussion dauerte dennoch an, weil sie mit der Entscheidung kämpften, das Teilprojekt nicht zur Förderung zu empfehlen und damit aus dem SFB auszuschließen. Es stellt sich an dieser Stelle die Frage, warum die Gutachtenden, die das Projekt als sehr gut und besser einschätzten, sich nicht zu Wort gemeldet haben. Diese Frage wird im folgenden Abschnitt diskutiert.

3.3.1 Expertentum

Die Auswertung der Gutachtendeninterviews zeigt, dass generell die Gutachtenden, die das Projekt der Gruppe vorstellen, vom Rest der Gutachtenden als Expert/inn/en für das Sachgebiet des Teilantrags angesehen werden. Wenn sich die Expert/inn/en in ihrem Urteil einig sind, verlassen sich die anderen Gruppenmitglieder auf ihre Bewertung. Sie sind bereit, ihre eigene Einschätzung zu ändern oder zurückzuhalten, wenn sie von der Expert/inn/enmeinung abweicht, da sie davon ausgehen, dass die Vorstellenden das Teilprojekt besser beurteilen können.

Das folgende Beispiel verdeutlicht dies: Ein/e Gutacher/in hat sich beim Durchlesen eines Teilprojekts darüber gewundert, dass der/die Teilprojektleiter/in die letzten zehn Jahre nur eine Arbeit veröffentlicht hat. Er/sie bezeichnete diesen Sachverhalt als ein *„Desaster"*. Der Vortragende des Projekts erläuterte, dass diese veröffentlichte Arbeit des Antragstellers/der Antragstellerin von höchster Qualität wie Priorität für das Fachgebiet sei und es sich bei dem Antragstellenden um einen/eine exzellente/n Wissenschaftler/in handle. Dies führte dazu, dass der/die Gutachtende sich von dem Experten/der Expertin überzeugen ließ und seine/ihre Zweifel an der Publikationsleistung des Antragsstellenden in der Gruppendiskussion nicht weiter ausführte.

> *„Als ich dieses Projekt gesehen habe, ich kenne mich ja im Detail nicht aus, das ist ja fast nicht zu vertreten. Und dann sehen Sie aber, dass es trotzdem funktioniert, weil dann jemand, der ein Experte ist und halt sagt: Nein, vorsichtig, das ist eine ganz, ganz tiefe Arbeit. Er hat sich also darauf konzentriert, jahrelang daran gearbeitet. Und das ist aber ein absolutes Superergebnis. Und dann (…) das ist jetzt auch veröffentlicht und so weiter, also kann man es glauben, dann akzeptiert man das auch. Das ist also jetzt in dem Fall dann halt eben wirklich eine Expertenmeinung, die wirklich auf die besondere wissenschaftliche Qualität, auch die Ausnahmesituation, nicht in das normale Raster eigentlich reinpasst. (…) man [glaubt] demjenigen, der wirklich ein echter Experte ist, von dem man sagt, ich kann das wirklich beurteilen und ich finde, das ist wirklich Klasse, dann akzeptiert man das in der Regel." (Gutachter/in C)*

Eine Diskussion findet erst dann statt, wenn sich die beiden Vorstellenden widersprechen. Ist dies nicht der Fall, findet in der Regel keine Gruppendiskussion statt. Zu einem ähnlichen Ergebnis kommt Langfeldt (2002; 2004). Sie untersuchte sechs unterschiedliche Panels, die Evaluationen der norwegischen Forschungslandschaft vornahmen. Die Panels waren hinsichtlich ihres Aufbaus und organisatorischen Vorgehens unterschiedlich zusammengesetzt. Langfeldt (2002, 2004) kam zu dem Ergebnis, dass der Entscheidungsfindungsprozess durch eine eindeutige Aufgabenteilung gekennzeichnet war und dass wenig Interaktion zwischen den Gutachtenden stattfand. Lediglich in den Fällen, in denen es eine Überschneidung der Kompetenzbereiche von Gutachtenden gab, kam es zu Diskussionen über Bewertungen. Langfeldt (2004) schlussfolgert deshalb, dass überlappende Kompetenzbereiche zwischen Gutachtenden eine wichtige Vorraussetzung sind, damit es zu Diskussionen über wissenschaftliche Qualität kommt. Aufgrund ihrer Ergebnisse fordert sie: „Two experts assessing each object under review and some time for discussing the results would be the minimum needed if expert panel evaluations are to have some function exceeding individual review reports when it comes to assessing the quality of research" (Langfeldt 2004: 60).

3.3.2 Aktivität der Gutachtendengruppe

Die Auswertung der Beobachtungsprotokolle hinsichtlich der Aktivität der Gutachtendengruppe bestätigt das Ergebnis, dass keine Gruppendiskussion stattfindet, wenn die beiden Vorstellenden in ihrem Urteil übereinstimmen. Abbildung 4 verdeutlicht diesen Sachverhalt exemplarisch für die Diskussion einer SFB-Klausursitzung. Es ist abgetragen, wie oft sich die einzelnen Gutachtenden an der Diskussion beteiligten. Grau eingefärbte Felder markieren, welche Gutachtenden sich als Vortragende besonders auf das Teilprojekt vorbereitet haben. Zum Beispiel waren die vortragenden Expertinnen/Experten des Teilprojekts B2 beide der Meinung, dass es sich um ein exzellentes Projekt (Note 1) handele und keiner der anderen Gutachtenden hat diese Einschätzung durch einen Redebeitrag ergänzt.

Abbildung 4: Aktivität Gutachtergruppe

Gutachtende / Teilprojekte	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	Note
A1						5			2			2
A2		2				4	1		4			1
A3			1		1	5		1				2
A4		1			7	3			3			3
A5	3	6	4			7		2	3	2		3
B1	3					1	1		1		1	1
B2							2				1	1
B3						2	7		1		7	2
B4					3	7	2		1		3	3
B5	1	1				1	3	3				1
B6	2					1		2	1			1
B7	4	4				4	2	5	4			1
C1	krank	6	2	1		9	3	3	1	16		2
C2	krank	3								5		1
C3	krank	3						1		2		1
C4	krank	5	2			1	1	1		3		3
C5	krank		12	2		7	3			5		3
C6	krank	6	4	3		8	1			3		5

Abbildung 4 verdeutlicht darüber hinaus, dass es in einer Gutachtendengruppe immer Menschen gibt, die ruhiger sind und andere, die sich intensiv an der Diskussion beteiligen. Zum Beispiel hat der Gutachtende G6 insgesamt 70 Wortbeiträge, an zweiter Stelle folgt der Gutachtende G2 mit 41 Wortbeiträgen. Das sind rund 40 Prozent weniger. Am seltensten beteiligte sich die/der Gutachter/in G4 mit lediglich einem Beitrag außerhalb des ihm/ihr zugewiesenen Teilprojekte (in der Abb. grau unterlegt).

Die Tabelle verdeutlicht, dass es in einer Gruppe Personen mit verschiedener Beitragsaktivität gibt. Offen ist an dieser Stelle, welchen Einfluss die Personen, die viel sagen, auf die Entscheidungsfindung haben. In den Interviews unterscheiden die Gutachtenden zwischen Personen, deren Beiträge ein hohes Gewicht haben und solchen Personen mit einem starken Redebedürfnis.

Das folgende Zitat eines Gutachters/einer Gutachterin steht stellvertretend für andere Aussagen von Panelmitgliedern mit ähnlichem Inhalt:

> *„Es gibt in solchen Gremien immer ein bis zwei, maximal drei Platzhirsche (…), die natürlich das Meinungsbild insgesamt schon sehr prägen. (…) Ich glaube, dass die Meinung der Platzhirsche immer dominiert." (Gutachter/in D)*

Ein weitere/r Gutachter/in beschreibt das Phänomen, dass sich unter den Vielrednern und -rednerinnen im Laufe der Diskussion einige herauskristallisieren, die eine dominierende Rolle im Urteilsfindungsprozess übernehmen.

> *„Sie haben gesehen, es gibt immer Kollegen, die sich sehr wenig beteiligen oder sehr ruhig sind. Jeder ist anders. Im Laufe der Diskussion kristallisieren [sich] unter denen, die viel sagen immer ein, zwei oder drei heraus, die ein bisschen die Meinungsführerschaft haben." (Gutachter/in E)*

Wenn Kritikpunkte von einer eher ruhigeren Person geäußert würden, hätte diese es dann teilweise schwerer sich durchzusetzen, führt der Gutachter weiter aus.

4 Diskussion

Zusammenfassend kann festgehalten werden, dass die Benotung der Teilprojekte ganz wesentlich von dem Urteil der Vortragenden abhängt und keine Diskussion in der Gutachtendengruppe stattfindet, wenn zwischen den Expertenurteilen Übereinstimmung im Urteil herrscht. Es stellt sich vor diesem Hintergrund die Frage, ob der Aufwand, den eine Gruppenbegutachtung an Zeit und finanziellem Umfang mit sich bringt, gerechtfertigt ist – oder ob es nicht ausreichen würde, das Statement der vortragenden Expertinnen und Experten schriftlich einzuholen und in Zweifelsfällen weitere Gutachten anzufordern. Die Gutachterinnen und Gutachter sehen den finanziellen und zeitlichen Mehraufwand der Gruppenbegutachtung als gerechtfertigt an und ziehen diese Form der Begutachtung der Einzelbegutachtung vor. Sie schätzen den Prozess der Gruppenbegutachtung, zum einen wegen des Austausches mit Kolleg/inn/en und Antragstellenden sowie der Möglichkeit, ihre Einschätzung mit denen der anderen zu vergleichen, zum anderen deswegen, weil sie davon überzeugt sind, dass die Gutachtergruppe Fehleinschätzungen einzelner auffangen kann. Die folgenden Zitate veranschaulichen diese Einstellung gegenüber der Panelbegutachtung beispielhaft:

> *„Ja, das ist so ein Mosaikspiel, wo jeder seinen Baustein dazu gibt und dann wird es zu einem Bild. Und das kann man eben wesentlich besser machen, wenn man (…) zusammensitzt als ganze Gruppe und nicht sagen wir mal nur per Computer oder schriftlich." (Gutachter/in F)*

> *„Aber nichtsdestotrotz gibt es eben als Regulativ die Auseinandersetzung mit den Antragstellern, und es gibt (…) das direkte Regulativ in dem Gutachtergremium, wo ich (…) eine Rückmeldung auf meine eigene Einschätzung bekomme. Und die finde ich gut, also, ich mag den Austausch (…) mit Kollegen, und auch (…) die Urteilsfindung, und denke, dass es (…) einfach ein fairer Prozess ist, (…) der unter Umständen länger dauert natürlich, aber (…) ich glaube, das ist der Sache angemessen (…)." (Gutachter/in G)*

> *„Gruppe ist da besser. Ist viel besser. Man kann sich wesentlich besser abstimmen. Man kann auch mal irgendwas durchdiskutieren, wo einer sagt, das sollte man vielleicht doch nicht machen, und man sollte andere Aspekte reinbringen. Oder man sollte diesem oder jenem die Vorhand lassen in irgendeinem Teilprojekt. Und das kann man in einer Gruppe wesentlich besser durchdiskutieren. Also irgendwo flexibler, fairer." (Gutachter/in H)*

Die Aussage, dass die Gutachtenden besonders den Austausch mit den Kolleg/inn/en an Gruppenbegutachtungen schätzen, erscheint nur auf den ersten Blick als Widerspruch zu der Beobachtung, dass in vielen Fällen keine Diskussion stattfindet. Eine Gruppendiskussion findet ganz besonders dann statt, wenn die Expert/inn/enmeinungen voneinander abweichen. In diesem Moment bringen sich andere Gutachtende mit ihrer Meinung in die Diskussion ein. Das heißt, es herrscht eine ergebnisorientierte Atmosphäre, die dafür sensibilisiert ist, wo Debatten notwendig sind. Nicht alle Diskurse werden ausgeführt, aber wo es notwendig erscheint (z.B. bei Dissens) wird davon Gebrauch gemacht und diskutiert. Die Gutachtenden empfinden das Verfahren dabei mehrheitlich als fair und sehen auch für sich die Möglichkeit, ihr Urteil kritisch zu hinterfragen.

> *„Ich muss sagen, dass in der Regel diese SFB-Begutachtung des einzelnen Projekts extrem fair ablaufen. Was man bei Einzelanträgen, wo das ja anonym ist, nicht immer so ist. Aber hier in diesen Gruppenanträgen, da spielen ja manchmal auch ganz interessante gruppendynamische Dinge eine Rolle, die da so abgehen. Das finde ich gut, dass es diese Korrektiven gibt. Ich sehe mich da auch selbst manchmal sehr wohl korrigiert, dass ich vielleicht ein Projekt relativ, sagen wir mal, negativ sehe. Und mir die Argumente von anderen Kollegen anhöre und sage dann okay, kann man auch so sehen. Und dann vielleicht zu einem milderen Urteil komme. Ich denke, dass diese Ebene ganz wichtig ist und je häufiger man das macht, eine umso größere Rolle spielt das." (Gutachter/in I)*

Die hier vorgestellten Ergebnisse haben vorläufigen Charakter, sie deuten aber zusammenfassend darauf hin, dass Effekte wie Gruppenpolarisierung und Meinungsführerschaft bei Panelbegutachtungen auftreten und das Begutachtungsergebnis beeinflussen. Inwiefern sich diese Ergebnisse auch in der Hauptphase des Projekts bestätigen, werden die weiteren Auswertungen zeigen.

Literatur

Alexiades, M. N. / J. W. Sheldon (eds.), 1998: Selected guidelines for ethnobotanical research: A field manual. 2. print. Advances in economic botany 10. Bronx: N.Y. Scientific Publ. Depart.

Bernard, H. R., 2002: Research methods in anthropology: Qualitative and quantitative approaches. 3rd ed. Walnut Creek, CA: AltaMira Press.
Online: http://www.gbv.de/dms/bowker/toc/9780759101470.pdf.

Borgatti, S. P., 1994: Cultural Domain Analysis. Journal of Quantitative Anthropology 4(4), 261-78.

Borgatti, S. P., 1996: Anthropac. Natick, MA: Analytic Technologies.

Burnstein, E. / Vinokur, A., 1977: Persuasive argumentation and social comparison as determinants of attitude polarization. Journal of Experimental Social Psychology 13(4), 315-32.

Daniel, H. / Mittag, S. / Bornmann, L., 2007: The potential and problems of peer evaluation in higher education and research, in: *Alessandro Cavalli (ed.):* Quality assessment for higher education in Europe. London: Portland Press, 71-82.

Dennis, A. R., 1996: Information exchange and use in small group decision making. Small Group Research 27(4), 532-50.

DeVries, D. R. / Marschall, E. A. / Stein, R. A., 2009: Exploring the peer review process: what is it, does it work, and can it be improved. Fisheries 34(6), 270-79.

Goethals, G. R. / Zanna, M. P. , 1979: The role of social comparison in choice shifts. Journal of Personality and Social Psychology 37(9), 1469-76.

Isenberg, D. J., 1986: Group Polarization: A Critical Review and Meta-Analysis. Journal of Personality and Social Psychology 50(6), 1141-51.

Johnson, V. E., 2008: Statistical analysis of the National Institutes of Health peer review system. PNAS - Proceedings of the National Academy of Sciences of the United States of America 105(32), 11076-80.

Köckeis-Stangl, E, 1982: Methoden der Sozialisationsforschung, in: *Hurrelmann, K. / Ulrich, D. (Hg.):* Handbuch der Sozialisationsforschung. 2. Aufl. Weinheim [u.a.]: Beltz, 321-70.

Lamont, M., 2009: How professors think: Inside the curious world of academic judgment. Cambridge Mass: Harvard University Press.

Langfeldt, L., 2001: The decision-making constraints and processes of grant peer review, and their effects on the review outcome. Social Studies of Science 31(6), 820-41.

Langfeldt, L., 2002: Decision-making in expert panels evaluating research: Constraints, processes and bias. Oslo: University of Oslo.

Langfeldt, L., 2004. Expert panels evaluating research: Decision-making and sources of bias. Research Evaluation 13(1), 51-62.

Larson, J. R. / Foster-Fishman, P. G. / Keys, C. B., 1994: Discussion of shared and unshared information in decision-making groups. Journal of Personality and Social Psychology 67(3), 446-61.

Moscovici, S. / Zavalloni, M., 1969: The group as a polarizer of attitudes. Journal of Personality and Social Psychology 12(2):125-35.

Myers, D. G., 1978: Polarizing effects of social comparison. Journal of Experimental Social Psychology 14(6), 554–63.

Nijstad, B. A., 2009: Group performance. 1. publ. Social Psychology. Hove: Psychology Press.
Online: http://www.gbv.de/dms/bowker/toc/9781841696690.pdf.

Obrecht, M. / Tibelius, K. / D'Aloisio G., 2007: Examining the value added by committee discussion in the review of applications for research awards. Research Evaluation 16(2), 79-91.

Olbrecht, M., / Bornmann, L., 2010: Panel Peer Review of Grant Applications: What Do We Know from Research in Social Psychology on Judgment and Decision Making in Groups? Research Evaluation 19 (4), 293-304.

Overbeke, J. / Wager, E., 2003: The state of the evidence: What we know and what we don't know about journal peer review, in: *Godlee, F. / Jefferson, T. (eds.)*: Peer review in health sciences. Medical research. London: BMJ Books, 45-61.

Sanders, G. S. / Baron, R. S., 1977: Is social comparison irrelevant for producing choice shifts? Journal of Experimental Social Psychology 13(4), 303-14.
Schnegg, M. / Lang, H., 2008: Die Analyse kultureller Domänen: Eine praxisorientierte Einführung. Methoden der Ethnographie(3).
Schulz-Hardt, S. / Brodbeck, F. C. / Mojzisch, A. / Kerschreiter, R. / Frey, D., 2006: Group Decision Making in Hidden Profile Situations: Dissent as a Facilitator for Decision Quality. Journal of Personality and Social Psychology 91(6), 1080-93.
Weller, A. C., 2001: Editorial peer review: Its strengths and weaknesses. Medford N.J. Information Today.
Wessely, S., 1998: Peer review of grant applications: What do we know? The Lancet, July 25.

Michèle Lamont

Pragmatic Fairness: production of the sacred while observing the rules

In my book *How Professors Think*[1] I study the way in which members of committees responsible for assessing requests for research grants go about separating the wheat from the chaff. What cultural references do they use to assess these research projects? How do they consider the assessment process? Where, in their opinion, is excellence to be found? Is it in the project itself? Or is it the result of interaction between the assessors? Moreover, does assessment culture differ from one discipline to another? Today, I shall restrict myself to answering only one of the many questions raised in my book, the question of why the members of a committee believe in the fairness of their deliberations. This analysis is based on 81 interviews carried out with assessors and programme officers linked to 12 multi-disciplinary committees tasked with the assessment of projects. These committees award grants and subsidies in Social and Human Sciences to researchers and universities in the United States. For two years I have been able to study the deliberations of decision-making committees of five foundations or prestigious national committees which provide financial support for the work of Ph.D. students or of full and untenured professors. These include the *Social Science Research Council*, the *American Council for Learned Societies*, the *Woodrow Wilson National Fellowship Foundation*, the *Society of Fellows* of an elite American university and one of the principal American foundations financing research in the Social Sciences. Further details on the collection of data and on the programmes which I studied can be found in my book, *How Professors Think*.

One of the pieces of evidence which emerges from the interviews that I carried out with members of committees of experts is their conviction that, just as the cream generally rises «naturally» to the surface of the milk, they are capable, in most cases, of identifying the best applications - of separating the wheat from the chaff - although the margin of choice is often very slight. And they consider that, in general, the assessment processes work reasonably well. The question which I shall put today is therefore this: with which rules must the assessors comply to be able to believe that they will succeed jointly in making fair decisions? To define the scope of these rules, I have questioned the members of these varied assessment committees about their past experience - particularly about «normal» cases as well as about the extreme cases which they have come across. I asked them to describe to me the best and the worst committees of which they have been members and to explain to me on what they base their (negative or positive) opinion of these committees. I also asked them about which assessors they consider to be the best and the worst and about cases where the assessment rules were not (in their opinion) observed, in order to be able to identify the rules made obvious by the disputes. In this analysis, I am not trying to discredit the process, but rather to understand what conditions make the process possible and ensure its success and its social achievement.

Once more, with a few exceptions, the committee members whom I have met consider that their deliberations are fair and that the assessors succeed all in all in picking out the most deserving projects. In general, they believe that it is excellence which governs the selection process. The fact that they are all immersed in a research culture guided explicitly by the pursuit of excellence constitutes the touchstone which prevents them from imagining that the result of the deliberations could be the expression of an «old boy network» or of favouritism. On the contrary, they are happy to abide by implicit common rules, which, they believe, help them to pick out the best projects, as defined by the culture to which they belong.

Max Weber (1858-1917) and Emile Durkheim (1864-1920) wrote about the production of belief (Weber 1971; Durkheim 1912). In *Économie et société*, Weber was especially interested in the

1 Lamont, M., 2009: *How Professors Think*. Cambridge, Massachusetts/London, England: Harvard University Press.

production of legitimacy, suggesting that legal-rational legitimacy demands the use of impersonal, abstract and consistent rules, which requires that personal interests are set aside. As for Durkheim, he has written about the production of religious feelings and about the mechanisms by which one comes to be involved in the sacred. In *Les formes élémentaires de la vie religieuse*, he suggests that the sacred is defined by its separation from the profane thanks to ritual.

We will see that the work of the members of assessment committees matches the procedures described by Weber and Durkheim. It is possible, moreover, to isolate a certain number of rules followed by the members of these committees, rules which lead them to believe in the legitimacy of the fruit of their deliberations. Some of these rule have the aim of standardising the procedures by seeing to it that personal interests are put aside while other rules are aimed at separating the sacred (the identification of excellence) from the profane (that is to say, personal interest, idiosyncratic preferences and everything which smacks of favouritism)[2].

I will thus describe the customary rules by virtue of which the members of committees come to believe in the efficacy of their collective assessment. We will see that the members of these committees follow implicitly the rules of joint deliberations, which consist in taking the best decision possible by having the benefit of someone else's opinion, by virtue of the confidence given to the skills of specialists who are familiar with the discipline to which the project is relevant. Since assessment of the different projects depends on the context of this assessment, the experts can vote strategically, without in any way losing their faith in the fairness of the process.

1 The production of legitimacy and of the sacred

Collective faith in the legitimacy of the fruit of the deliberations is the result of the assessors' acceptance of rules which have never been formalised nor made plain by the president of the committee on which they sit. In their capacity as university graduates, they have become familiar with these rules in the course of their professional careers, especially when they have found themselves on committees recruiting colleagues or students applying for admission to doctoral programmes.

The use of consistent, universal criteria is essential for the legitimacy of the procedure, which is to say that the members of the committees, by selecting projects which will receive financial support are duty-bound not to take into account their personal ties to applications or idiosyncratic preferences. They have to display credibility and competence in the eyes of their colleagues, so as to be able to convince them of the value of research projects which they regard very highly. Their ability to show themselves ready to listen to others is important as much for creating their legitimacy as an assessor as for their expertise as a university researcher. In fact, it is this ability on which is founded a reputation for collegial reputation and wisdom. These qualities are clearly demonstrated in the descriptions in the interviews of a «good» or «bad» assessor.

2 I consider these rules as being «customary rules» because they are often tacit, informal and learned on the job, as the researchers are confronted with the production and assessment of the research in the course of their doctoral studies and of their professional socialisation.

2 Characteristics of a «good» committee member

The qualities which characterise a « good » committee member are the following:

2.1 Being perfectly prepared and ready to discuss the applications

The preparation of the application and the mastery of details make it possible to consider the case with full knowledge of the facts and to advance arguments which will persuade the other members of the team. Preparation makes it possible for evaluators to be able to respond to the - unpredictable - reactions of other experts. Showing an acute sense of responsibility and of seriousness in the work is the foremost quality sought in a member of a committee.

2.2 Showing an open and experienced intellectual spirit

It is expected that an expert will master a major part of the literature relevant to his discipline and that he will have a good passive knowledge of the many fields which are related to it, so that he can rapidly pick out the strengths and weaknesses in the applications to be discussed in these disciplines.

2.3 Being concise

It is essential not to waste time in talking too much or too slowly or by entering into unnecessary detail. This is because committees must accomplish a large quantity of work in very little time. Thus, for a historian whom I interviewed, his worst experience was of a committee where an anthropologist spent a great deal of time discussing applications which he had not even read, and in giving his opinions on everything and nothing.

2.4 Expressing oneself so as to be comprehensible despite differences between disciplines

This talent is essential if one wishes to be able to explain to the assessors the strengths and weak-nesses of the projects to be assessed. This skill is particularly valuable on multi-disciplinary committees, where the members are dependent on each other's knowledge and skill, for they are called upon to give opinions on work which is aimed at enriching research fields with which they have little familiarity.

2.5 Respecting the skills and opinions of others

A historian explained to me why he always felt at ease with two persons who were part of a committee on which he sat: «I liked the way they behaved towards the other assessors; they always showed respect. I felt at ease with them and knew that everything I might say would be listened to with respect».

Feeling valued helps one to listen more attentively to others and to be open to their opinions. The ultimate sign of respect is to show that one is prepared to listen to the opinions of others and to acknowledge their understanding of their field.

Intellectual breath, not just in a particular specialism, and common sense and being open to others are characteristics of a «good» assessor. These characteristics differ from those which figure in most of the literature devoted to peer review. In that literature emphasis is laid on cognitive aspects only. As I explain in my book, *How Professors Think*, I see assessment as an emotional, moral and social process, as opposed to the more classical approach which emphasises the intellectual/scientific and cognitive aspects of assessment.

3 Democratic deliberation: making a territory one's own, recognising skills, respecting the sovereignty of different disciplines

In the wake of Jürgen Habermas, political theorists have written reams about the rules which govern democratic deliberations and the criteria according to which the deliberation should be judged. These conditions involve reciprocity and mutual respect as well as transparency. Each person taking part in a democratic deliberation must be able to exercise complete freedom of expression and his vote must be equally weighted against those of the others; the arguments put forward must be rational and the deliberation must be consensual and centred on the common good.

In many ways the deliberations of evaluation committees follow principles analogous to those in democratic deliberations. Reciprocity and the principles of equality and relevance to the achievement of the common good are considered important factors in the process. What is more, the assessors are expected to convince each other using arguments of reason, and to observe the principles of democratic decision-making. Since their aim is to take decisions, which are both specific and informed, each member must be free to express himself without fear of reprisals and s/he must be given the opportunity to be listened to as much as any other member of the committee. The meeting is deliberative to the extent that it allows the sharing and the confrontation of differing opinions.

In practice, these ideal conditions are limited by the composition of committees. The members differ in age, race and gender and they come from institutions which do not enjoy equal prestige. Although, according to the protocol, all the members are equal, these characteristics of age, gender and race have an effect on how the opinions of each member are received and have repercussions in taking decisions. Moreover, there are considerable variations in the level of expertise of the committee members, which varies with the subjects and disciplines represented in the projects under discussion. Thus, as Weber's analysis of the construction of legal-rational legitimacy would have predicted, the opinion of this colleague or that colleague carries more or less weight depending on his degree of expertise as regards the object of the deliberation or the discipline of the candidate. Moreover, the assessors sometimes prefer not to exercise their right of expression when they do not feel able to do so with competence. This is what I call the rule of *respect for the sovereignty of disciplines*. This is a primary customary rule which structures the deliberations of committees of experts which I have studied.

A large part of the work of a deliberation consists in debating what is the correct label to attach to any given project. The assessors try to convince each other that a project has an precise plan (without being too restrictive), that it is ambitious (without being too risky), or that it is up to date (without being too «trendy») and that it has been carefully targeted (and is not desperately vague). The ability to attach a label generally depends on the fact that an assessor can demonstrate to the other committee members that his prior experience qualifies him to give a particularly well-informed judgement.

This is what establishes the second customary rule, which is consistent with the first rule: *that which consists in deferring to established expertise*. This rule is all the more important for it is essential to be able to have confidence in the judgement of others if a committee hopes to be able to compare efficiently numerous project proposals, each as new as it is different from the others. This is illustrated by a professor of English who reports:

> *I tended to give a good assessment to a project proposal which seemed exciting, but which came from a field with which I was not very familiar, until another member of the committee intervened by saying, «This project is not innovative, you know!» How could I have been aware of that?*

The most acrimonious disputes arise when the assessors think that their colleagues are not respecting their competence to assess the projects for which they consider themselves most qualified. Or else

when several assessors think themselves equally qualified to assess a project. This was the case in a project which some judged to have an out-dated framework. Thus, one historian explained that another historian

> *... did not have the right to try to make us accept a project which was not well conceived. She wanted to persuade us that, even if the project plan was not perfectly well constructed, its author was someone very good and that it was worthwhile to consider the application carefully. Of course it is possible to abandon the criterion of competence, but thereafter, one still has to succeed in convincing two other persons that it is a good idea.*

Since, most often, one submits oneself to the rules of deference to competence and to respecting the sovereignty of disciplines, the exchanges of opinion between the members of a committee usually take place amicably. This amiability is seen as the proof that a committee works well, even if the culture of deference somewhat limits the desire of the members to enter into animated discussions on the merits of project plans submitted for their approval.

4 Forming alliances, strategic voting and exchanging favours

Weber would say that legal-rational legitimacy depends on the fact that universal, consistent criteria are applied to all decisions. Universality and consistency in the assessment criteria is a third customary rule particularly dear to the assessors. In theory, the application of this rule gives all projects an equal chance to receive finance - all things being equal.

This rule is often broken, which is evident when the assessors describe the formation of alliances, strategic voting and the system of «you scratch my back and I'll scratch yours» (or exchange of favours). They generally take part in forming alliances, in strategic voting and "horse-trading". However they consider this behaviour as compatible with universalism or with the capacity of the assessors to separate the wheat from the chaff.

In line with the spirit of a democratic deliberation, which requires that all the participants contribute to the assessment equally and without being influenced by external factors, many persons whom I interviewed stated that they had never witnessed nor had taken part in strong alliances between different members of a committee, even if, from time to time, they felt more in agreement with some members than with others. If they exist, the assessors do not consider that these personal and intellectual affinities are corrupt or illegitimate. They cannot do anything else than judge projects using the intellectual tools available to them, which, inevitably, coincide more or less with those employed by some of the other members. Having regard to the use of universal and consistent assessment criteria deemed to characterise the joint task, it is not surprising that the members of committees minimise the role of strategic voting and reciprocal favours.

Strategic voting consists in giving a bad assessment to one project, so that other projects will have a better chance of emerging victorious from the deliberation. This also consists in overrating mediocre or disputed projects to improve their chances of being subsidised.

As for exchanging favours, this consists in acting so that the wishes of other members are carried out, in order to increase the probability that they will do as much for you when the time comes.

These two practices are accepted and are even considered normal, although some committee members think that they do not always allow the intrinsic qualities of each project to be assessed as they deserve.

In fact, strategic voting is considered part of the normal course of things. For example, an assessor told me she considered it normal to vote strategically in favour of a project, in view of the intrinsic qualities of the project. She admitted having given a more positive mark to one project to guarantee that it would be discussed, for she expected that other members of the panel might not like it. What is openly condemned is less overrating a project than giving a lower mark than a project

deserves - which is considered, to some extent, as «playing politics».

In any case, it is difficult to find a distinction between normal voting and strategic voting, because all voting is strategic insofar as it is attempting to support or prevent the financing of a project. However, voting is not considered strategic unless it takes into account not only the merits of the project in question, but also the desire to support or hinder the behaviour or the voting of another assessor.

Strategic voting and exchanging favours become vitally important towards the conclusion of deliberations. It is when the final grants are being allocated that exchange of favours becomes most obvious. This is the moment when factors alien to the true value of the project come into play and when the assessors engage in «you scratch my back and I'll scratch yours» which would have been unthinkable until this point. This was especially the case for an assessor who explained that he had supported a project he did not like because he was convinced that a project, of which he was in favour was going to be held back.

The majority of members of committees agree that strategic voting and exchanging favours are legitimate and inevitable characteristics of the procedure in which they take part. The dynamics of ranking the applications is such that the majority of judgements are made in a relational and contextual fashion, by comparing sub-groups of applications. It is in this context that the assessors come to think, strategically, about what they could reasonably do to influence the ranking, in view of the fact that all the assessors want to convince their colleagues that their judgement is well founded. One historian expressed it thus: «You win some and you lose some; you cannot always have your way».

Moreover, the members of committees are aware that the excellence of one project is determined in comparison with that of other projects that the field of comparison is made up of all the projects which are to be judged and that contextual ranking is at the heart of the art of assessment.

5 Renouncing personal interest and personal relationships ('old boy' network, populism and particularism)

In contrast to strategic voting and exchanging favours it is thought that the influence of personal interest and contacts on the fruits of deliberations is completely illegitimate because it corrupts the process and is antithetical to the principle of using the universals and consistent criteria referred to above. When I asked one assessor what his reaction would be if one of the committee members declared, «this is a student of one of my close colleagues. I would really like his project to be funded», he replied:

> *This could not become a factor. It is unthinkable. Of course, the members of a committee all have leanings towards one project or another, but they try to ignore these leanings.*

He added that specialists in any given field are more likely to make the financing of a project relevant to this field more difficult:

> *The more specialists on the Middle East there are on the committee, the fewer are the projects relevant to the Middle East which will receive finance. This is because people are more demanding towards that which they know, to the point that one is sometimes obliged to find ways to soften their views a little.*

Thus, following Durkheim, it can be said that personal interest is impure and that it goes against excellence, which is considered sacred, which is deemed to guide assessment and which is, in theory, protected by the customary rules aimed at limiting corruption.

The funding organisations generally give very explicit instructions concerning when evaluators should abstain from discussion pertaining to the research of a close friend, a colleague or a student.

Although they are by no means obliged to do so, the members of a committee often volunteer the information that they are acquainted indirectly with this or that person connected with the project which is of being assessed. They will say for instance «The mentor of this student is a close colleague of mine» or else, «I know this student's academic director very well and I have every confidence in the content of his letter». Divulging such connections forms part of the customary rules that contribute to the confidence in the fairness of the deliberations.

When I asked one of the assessors whether he considered that his work took on a sacred value, he replied:

> *Yes, I think that it protects a value which transcends the 'old boy' network and that sort of thing. And I would agree that what we are trying to do is something which is free from personal interest and limited points of view.*

Of course, it is almost impossible to completely eliminate the 'old boy' network and particularism from the assessment procedure, especially when academic competence and the assessment criteria are quasi «grafted» on to personal connections which the producers of knowledge maintain with each other.

In theory, the assessors must not be influenced by their social links. They must act as «independent agents», with no personal interest, so that particularism has no influence on the way in which decisions are taken. This is also the reason why the preferences of members of a committee tend to be expressed in universalist terms. But, in reality, these assessors cannot entirely cut themselves off from their involvement in academic networks. Their assessment criteria form part of a social context, just as do their networks and their contributions to their discipline.

6 Managing idiosyncrasies: openness of spirit and cognitive contextualisation

Collective faith in the fairness of a deliberation is maintained if the members make an effort to put aside their idiosyncratic tastes and preferences. However, it is very difficult to ignore these latter which, in fact, form an integral part of the assessment. Thus, it is thought that assessors tend to define the originality of a project with regard to the type of originality shown in their own work. They also tend to view favourably and to like that which accords best with their own research interests. Thus, one assessor recognises that «excellence is what is most like you». To minimise this tendency, he suggests that judgements based on taste should be distinguished from judgements based on expertise.

The fact that assessors do not formulate their preferences in terms of personal interest reveals the codes which they use to attempt to give a sacred character to their actions. They freely admit that liking and viewing favourably that which is like you leads to homophilic reproduction. At the same time, it is difficult, if not impossible, to imagine an assessment procedure which would totally suppress all personal preferences.

One anthropologist, fully aware of the need to avoid idiosyncratic judgements, summarised the problem as follows:

> *It is never possible totally to free oneself of one's own interest, one's own point of view, etc. These predispositions influence the decisions one takes. I don't know if it is necessarily a bad thing when the composition of a committee allows different points of view to be expressed and represented. There are people from different disciplines, who have different skills, who come from completely different types of university, and then there are individuals whose characteristics are obvious, such as those of race, ethnicity and gender. This is why I think that it is important to take into account these differences when the committee is being formed, because it is never possible totally to free oneself of one's own point of view, as is well known.*

The answers to the question, «What characterises a good committee member?» prove to what extent openness of spirit and tolerance of differences are prized. Extrapolation from the customary rules demonstrates that members of committees must show themselves to be pluralists in their methods of assessment, and that they are eager to assess the applications according to the epistemological and methodological criteria appropriate to the discipline of those who have submitted these applications. This is what my colleagues and I have called cognitive contextualisation (Mallard/ Lamont/Guetzkow 2009). This constitutes another important customary rule.

In describing the dynamics of his group, one of the judges summarises this principle thus:

> *One notices differences between people who work on fairly large data sets and who do quantitative research and then, on the other hand, those who do in-depth ethnographic studies in anthropology. Their methodologies are so different that it is difficult to say whether a general criterion applicable to both cases exists. Fortunately, I think that we were all ready to comprehend the projects in the terms appropriate for them and not to try to apply a more general criterion to them.*

Committees must, therefore, not challenge methodological or disciplinary traditions. The rules of the game demand that the principle of equality between the different methodologies be recognised, at least while the deliberations are taking place. In theory, the members must hold themselves back from emphasising too strongly the qualities of one discipline as opposed to another and from trying to bolster their own prestige or that of their discipline.

The value attributed to cognitive contextualisation serves to counterbalance the idiosyncratic tastes of the committee members; it encourages them to suppress personal tastes and to judge projects according to the reference criteria of the field from which the proposal comes. The assessors prize this diversity because it increases their feeling that the work of the committee produces a high-quality result - a result which reflects the intrinsic qualities of the project under consideration. For example, one historian, who has a penchant for hermeneutics, admitted that he was pleased when a political scientist stated that:

> *Every committee needs at least one person who is rigorous, empiricist, scientific, a specialist in the social sciences, who sees himself as such and who will be able to say why his criteria are as they are.*

One historian explained this attitude thus:

> *People tended to show a certain spirit of openness (by that I mean that we made the effort to keep an open and appreciative attitude towards that which is considered good work in fields other than our own). You must show yourself capable of « projecting yourself », of being able to imagine a type of work which is very different from our own, to acknowledge the criteria relevant to assess the other as compared with your own work. It seemed to me that everyone was entirely committed to this objective.*

These extracts from interviews show to what point methodological pluralism is essential for funding committees to work well. They also show how moral values and feelings are at the very heart of the assessment procedure, as opposed to being alien and corrupting, as some of the literature on assessment suggests.

7 Methodological pluralism: putting aside bias related to each discipline

If, in the course of a deliberation, the assessors keep themselves from emphasising the defects of other disciplines, this is the simple manifestation of another customary rule which consists in observing methodological pluralism, the sovereignty of disciplines, as well as the skills and opinions of other members of the committee. As is the case when one refrains from expressing doubts with regard to a methodology foreign to one's own discipline and from making disagreeable comments

on these disciplines - which would be perceived by the other assessors as extremely incorrect. This was well illustrated by one philosopher who criticised a historian for having described philosophy as a «sterile intellectual exercise» - an inexcusable faux pas in the context of a multi-disciplinary committee.

As we have already said, the legitimacy of the procedure depends, in part, on the members of committee who have fewer skills in a given discipline, deferring to those who do possess the requisite skills. A sociologist described as follows this dynamic in his description of the assessment of a controversial research project:

> *If a project plan concerning identity is involved, I will give it a fairly low assessment because it is not in line with the high-level criteria which apply in my own field. If then someone else - perhaps an anthropologist or a historian who will have greater competence in the field and who will be more familiar with the literature on identity - gives it a good assessment, what generally happens in these meetings is that the person who gave the project a bad assessment will listen very carefully to those who gave it a good assessment, especially if the latter comes from a field which gives them the relevant knowledge. Then, one makes concessions, a compromise must be reached.*

This same sociologist recognised that some of his assessments favour certain disciplines more than others. He expressed a desire to correct this bias. Such an attitude is honourable in the context of the customary rules which I have described. It is essential for the maintenance of belief that the deliberations are fair and finally confirms the importance of the rituals, which, according to Durkheim, are essential for the creation of the sacred. This attitude also protects the spirit of compromise which one historian described thus: «In general, I try to give the benefit of the doubt to projects coming from disciplines which I am not very familiar».

8 Conclusion

The preceding pages have neither the aim of demonstrating that the peer review system is a perfect system, nor of asserting that it is discredited. I have simply reviewed the customary rules followed by members of funding committees to persuade themselves that the procedure works well and that they will succeed, at the end of the day, in picking out the best projects. I have restricted myself to analysing the assessment conditions which lead members of these committees to believe in the legitimacy of their collective work.

In fact, in general, they think that peer review is an imperfect assessment procedure, but that one endeavours to make it as perfect as possible, while maintaining belief in the primacy of disciplines, by asserting the need to maintain a distance with regard to personal interests, and other customary rules, whose importance I have highlighted. This does not prevent this method of assessment from being accompanied by strategic voting, the exchange of favours and by references to idiosyncratic tastes. Thus, I demonstrate that peer review is a highly social and emotional procedure, while also being cognitive in the sense that the assessors invest in the process their sense of their personal honour and their dignity as experts. These constants, however, do not answer all the questions.

Thus, it may be that the assessment procedure analysed here applies only to committees composed of members from different disciplines. Perhaps uni-disciplinary committees are different - especially with regard to the rule about deferring to the experts or to those from the main disciplinary area. Doubtless, assessment of funding proposals may have different characteristics from those of assessment of scientific articles.

Perhaps also the customary rules analysed are made possible by the distinctive features of American higher education system, including its imposing size, institutional diversity, and geographic dispersion. To grasp the full impact of assessment on the customary rules would need, therefore, considerably more work. This is more particularly the case when one tries to analyse how these rules may apply

in a European context and to assessment committees which follow different procedures.[3]. This is what I have begun to do elsewhere. To follow...

Literature

Cousin B. / Lamont M., 2009: Les conditions de l'évaluation universitaire: quelques réflexions à partir du cas américain. Revue Mouvements, 60, 113-117.
Online: http://www.mouvements.info/spip.php?article409.

Durkheim E., 1912: Les formes élémentaires de la vie religieuse, Paris, PUF, 5è édition (1968), Collection Bibliothèque de philosophie contemporaine.
Online: http://dx.doi.org/doi:10.1522/cla.due.for2

Lamont M., 2009: How Professor Think: Inside the Curious World of Academic Judgment. Cambridge, Massachusetts/London, England: Harvard University Press.

Lamont M. / Huutoniemi K., 2011: Comparing Customary Rules of Fairness: Evaluative Practices in Various Types of Peer Review Panels, in: *Camic, C. / Gross, N. / Lamont, M. (eds.):* Social Knowledge in the Making. Chicago: University of Chicago Press, 209-232.

Mallard G. / Lamont M. / Guetzkow J., 2009: Fairness as Appropriateness: Epistemological Pluralism and Peer Review in the Social Sciences and the Humanities. Science, Technology and Human Values, 34 (5), 573-606.

Weber, M., 1971: Économie et société. Paris: Plon.

3 Cousin B. & Lamont M. (2009), « Les conditions de l'évaluation universitaire : quelques réflexions à partir du cas américain », *Revue Mouvements*, 60, pp. 113-117. Available on Internet http://www.mouvements.info/spip.php?article409; Lamont M. & Huutoniemi K. (2011), Comparing Customary Rules of Fairness: Evaluative Practices in Various Types of Peer Review Panels in C. Camic, N. Gross & M. Lamont (Eds.), 2011: *Social Knowledge in the Making*, Chicago, University of Chicago Press.

Eva Barlösius

Der Wandel der Ressortforschungseinrichtungen während des Evaluationsprozesses

Bis vor wenigen Jahren bildeten die Ressortforschungseinrichtungen, darunter werden die den Bundesministerien nachgeordneten Einrichtungen mit Forschungs- und Entwicklungsaufgaben zusammengefasst, eine „terra incognita" (vgl. Barlösius 2009; WR 2007).[1] Obwohl mit Politikberatung beauftragt und oft verantwortlich für hoheitliche Vollzüge, waren sie bis vor wenigen Jahren in der breiten Öffentlichkeit, aber auch im wissenschaftlichen Feld und in der Wissenschaftspolitik wenig bekannt. Damit ging einher, dass nicht spezifisch ausformuliert war, was diese Einrichtungen kennzeichnet, ob und wie sie sich von anderen Einrichtungen des wissenschaftlichen Feldes unterscheiden. Ebenso wenig bestand Klarheit darüber, wie sie sich im und zum wissenschaftlichen Feld positionieren. Besonders deutlich wird dies daran, dass

- es keine klare Charakterisierung der Forschung gab (betreiben sie Vorlauf- und Vorsorgeforschung, Anwendungsforschung, Grundlagenforschung, freie Forschung, oder repräsentiert Ressortforschung eine Forschung eigenen Typs?);
- undeutlich war, ob sie einen eigenen Nomos, ein eigenes „Grundgesetz" ausgebildet haben, auf welches sie sich als Alleinstellungsmerkmal berufen können;
- die Leistungskriterien, die für sie gelten und ihnen im staatlichen wie im wissenschaftlichen Feld Anerkennung verschaffen, nicht ausgearbeitet waren.

Dass diese und weitere Kennzeichen vielfach unbestimmt und außerhalb der Einrichtungen weitgehend unbekannt waren, bewahrte die Ressortforschungseinrichtungen, aber auch die zuständigen Ministerien davor, ihre Differenzen gegenüber und Übereinstimmungen mit anderen Einrichtungen im wissenschaftlichen Feld – wie den Universitäten oder den vier außeruniversitären Organisationen– zu bestimmen. Es enthob sie weiterhin davon, Stellung dazu zu beziehen, ob sie zum wissenschaftlichen oder staatlichen Feld oder zu beiden Feldern gleichermaßen gehören. Ebenso dispensierte es sie davon, die interne Strukturierung und Organisation der Einrichtungen zu begründen und ihre Spezifität zu charakterisieren, von der aus ein eigener Nomos beansprucht werden kann.[2]

Das änderte sich – nicht grundlegend, aber doch erheblich –, als der Wissenschaftsrat (WR) 2002 begann, die Ressortforschungseinrichtungen systematisch zu evaluieren. Die Evaluation durch den WR drängte die Einrichtungen der Ressortforschung dazu, sich zu positionieren und ihre Stellung zwischen Staat und Wissenschaft zu charakterisieren und gegenüber Dritten zu kommunizieren. Sie waren aufgerufen, sich als forschendes oder nicht selbst forschendes Institut zu präsentieren und zu verdeutlichen, was ihre Besonderheit ausmacht. Weiterhin hatten sie darzulegen, was aus ihrer Sicht gute Leistungen sind und wie sie intern ausgestaltet sein sollten, damit sie die Aufgaben der Ressortforschung erfüllen können. Auch die Ministerien waren aufgefordert, ihre Wissenschaftsbehörden darzustellen und nachzuweisen, warum ein privilegierter ministerieller Zugriff auf eigene Forschungseinrichtungen erforderlich ist. Ferner hatten sie zu begründen, was die dort erbrachten Leistungen von denen anderer forschender Institute unterscheidet und zu explizieren, warum sie diese institutionell und organisatorisch so eingerichtet haben. Der WR – als dritter Akteur –

1 Wichtige Studien zur Ressortforschung sind: Lundgreen u.a. 1986; Hohn/Schimank 1990; Schimank 2005; Döhler 2007; Krauss 1996.

2 Hinter diesen Aufzählungen verbergen sich keine außergewöhnlichen Ansprüche an derartige Einrichtungen; das Institut National de la Recherche Agronomique (INRA) hat beispielsweise eine wissenschaftsphilosophische und -soziologische Vortragsreihe organisiert, um sich Klarheit über diese Aspekte zu verschaffen (Bourdieu 1997; Chevassus-au-Louis 2007; Dejours 2003; Roqueplo 1997).

hatte sich kundig zu machen, was die Ressortforschungseinrichtungen charakterisiert und ob die von ihm üblicherweise herangezogenen Evaluationskriterien auch auf diese Institute anzuwenden seien. Obendrein hatte er eine wissenschaftspolitische Position gegenüber diesen Instituten zu entwickeln.

2002 setzte der WR eine Arbeitsgruppe „Rahmenbedingungen der Forschung in Ressortforschungseinrichtungen" ein, die sich auf die Ressortforschungseinrichtungen des Bundesministeriums für Ernährung, Landwirtschaft und Verbraucherschutz (BMELV) beschränkte und 2004 ihre Empfehlungen vorlegte (WR 2004). Im November 2003 beschloss der Haushaltsausschuss des Deutschen Bundestages, die Ressortforschungseinrichtungen auch der anderen Ministerien durch den WR evaluieren zu lassen. Im Mai 2004 stimmte der Deutsche Bundestag diesem Beschluss zu. Der Auftrag an den WR lautete, eine „aufgabenkritische Überprüfung der Ressortforschungseinrichtungen hinsichtlich der Notwendigkeit eigenständiger Forschung und deren wissenschaftlicher Qualität" vorzunehmen (WR 2007: 5, 34). Zu diesem Zweck setzte der WR 2004 wiederum eine Arbeitsgruppe ein, die zunächst 13 Ressortforschungseinrichtungen evaluierte. 2007 verabschiedete er übergeordnete „Empfehlungen zur Rolle und künftigen Entwicklung der Bundeseinrichtungen mit FuE-Aufgaben" (WR 2007). Noch vor der Verabschiedung dieser Empfehlungen bat die Bundesregierung den WR, die restlichen Einrichtungen zu evaluieren, was bis Anfang 2010 erfolgte. Im Herbst 2010 schließlich hat der WR eine zweite übergeordnete Stellungnahme zu den Ressortforschungseinrichtungen vorgelegt (WR 2010).

An dieser Stelle kann nur ein grober Überblick über diesen sehr umfangreichen und sich über acht Jahre hinziehenden Evaluationsprozess gegeben werden. Manche Aspekte müssen leider gänzlich vernachlässigt bleiben, etwa die aufschlussreiche Unterschiedlichkeit des ministeriellen Umgangs mit der Evaluation, andere können angerissen werden, beispielsweise die große Heterogenität der Ressortforschungseinrichtungen. Im Zentrum steht, zu zeigen, dass die Evaluation unterschiedliche Ebenen im Blick hatte und was sich jeweils während des Zeitraums der Evaluation änderte.[3] Für die Analyse baue ich auf meinen Studien zur Ressortforschung auf, in denen ich – orientiert an den Feldanalysen von Pierre Bourdieu – drei Ebenen unterscheide (Barlösius 2009).

1. *Position der Ressortforschungseinrichtungen im wissenschaftlichen Feld oder/und im staatlichen Feld.*
   Diese dokumentiert sich insbesondere in: a) der Zugehörigkeit (und der Zusammenlegung) von Einrichtungen, b) den Benennungen der Einrichtungen, c) der Klassifikation der Leistungen (eigene Forschung als Beleg der Zugehörigkeit zum wissenschaftlichen Feld vs. nichtwissenschaftliche Leistungen wie hoheitliche Aufgaben, Politikberatung, Normung etc. als Begründung der Zugehörigkeit zum staatlichen Feld).

2. *Interne Strukturierung der Ressortforschungseinrichtungen im Hinblick auf institutionelle Möglichkeiten zur Selbstkontrolle und -ausrichtung.*
   Dies kann analysiert werden anhand: a) der Orientierung an wissenschaftstypischen Institutionen der Selbstkontrolle, b) der Übernahme rechtlicher und bürokratischer Vorgaben, die typisch für das wissenschaftliche Feld sind, c) der wissenschaftlichen Selbstorganisation.

3. *Wissenschaftlicher Nomos, der ein eigenes „Grundgesetz" konstituiert und eine spezifische Logik eines Feldes begründet.*
   Hierfür können herangezogen werden: a) der Forschungstypus, b) ein spezifisches „principe d'unité" (Bourdieu 1997) der Ressortforschungseinrichtungen, das als Alleinstellungsmerkmal anerkannt wird.

Aus meinen Studien zur Evaluation leite ich die Hypothese her, dass Evaluationen wissenschafts-

3 Hier stellt sich das in der Soziologie bekannte Problem der Zurechenbarkeit. Einige Veränderungen waren bereits vorher geplant, von anderen ist dies anzunehmen, und weitere sind unmittelbare Reaktionen auf die Evaluation. Soweit ich eine Zurechenbarkeit herstellen kann, identifiziere ich sie, ansonsten enthalte ich mich einer solchen. Keineswegs beanspruche ich, die empirischen Möglichkeiten, eine Zurechenbarkeit zu bestimmen, vollkommen erschlossen zu haben. Dazu müssten weitere Quellen herangezogen werden.

intern in zwei Richtungen wirken: Zum einen stärken sie den „wissenschaftlichen Habitus“ und die darin enthaltene „maîtrise practique“. Zum anderen verhelfen Evaluationen zu einer einheitlicheren Auffassung und Beschreibung des wissenschaftlichen Feldes. Auf diese Weise tragen sie dazu bei, Übereinstimmungen und Differenzen innerhalb des wissenschaftlichen Feldes und gegenüber anderen Feldern deutlicher und strikter zu markieren (vgl. Barlösius 2006, 2007). Da es sich bei der Evaluation der Ressortforschungseinrichtungen um eine sogenannte Systemevaluation handelte (die Evaluation eines Feldes), beschränke ich mich auf den zweiten Aspekt. Entsprechend ist zu fragen, ob, veranlasst durch die Evaluation, einige der vorne aufgezählten Unklarheiten aufgehellt, die Übereinstimmungen und Differenzen klarer markiert und eine eindeutigere Positionierung vorgenommen wurde. Eine ähnliche Absicht verfolgte das BMBF mit der Evaluation. In seinem Bericht an das Bundesministerium der Finanzen (BMF) Nr. 99/04 kündigte es an: „Die Evaluierung der Ressortforschung durch den WR wird dazu beitragen, den Begriff der Ressortforschungseinrichtung näher zu erläutern und sein Profil zu schärfen.“ (BMBF 2004: 3)

1 Position der Ressortforschungseinrichtungen

1.1 Zugehörigkeit und Zusammenlegung der Einrichtungen

Die Zugehörigkeit der Institute zur Gruppe der Ressortforschungseinrichtungen wird formal darüber hergestellt, dass die Bundesministerien diese auf die entsprechende Liste des Bundesberichts Forschung setzen lassen.[4] Im Bundesbericht Forschung von 2004 waren 53 Einrichtungen verzeichnet und in dem aus dem Jahr 2006 51 Institute. In dem Bericht von 2008, der nach den ersten systematischen Empfehlungen des WR verfasst wurde, sind 46 Einrichtungen aufgeführt (BMBF 2008: 146-158). Diese Zahlen zeigen, dass es Veränderungen gegeben hat; allerdings stehen nur einige davon in direktem Zusammenhang mit der Evaluation durch den WR.[5] Die größte Veränderung ergab sich im Jahr 2007 durch die Neuordnung der Ressortforschungseinrichtungen im Geschäftsbereich des BMELV.[6] Sie war bereits vor der Evaluierung durch den WR geplant, allerdings wird im Gesetz an vielen Stellen auf Empfehlungen des WR verwiesen.

Der WR hat sich weitgehend an der im Bundesbericht Forschung von 2004 abgedruckten Liste von Ressortforschungseinrichtungen orientiert, sie nicht grundsätzlich hinterfragt und als Arbeitsgrundlage übernommen. Nur in ganz wenigen Fällen hat er empfohlen, Einrichtungen von der Liste zu nehmen, weil sie nicht, zu wenig oder nicht angemessen forschen. Häufig hat er dagegen gefordert, dass die Institute mehr eigene Forschung betreiben sollten, um ihre Aufgaben besser erfüllen zu können und ihrem Namen als Ressort*forschungs*einrichtungen gerecht zu werden.

Bemerkenswerter ist jedoch, dass im Bundesbericht Forschung 2008, wie erstmalig ein Jahr zuvor im „Konzept einer modernen Ressortforschung“ der Bundesregierung (BMBF 2007b), die Ressortforschungsinstitute, die dort bislang allesamt als Einrichtungen mit FuE-Aufgaben tituliert waren, in zwei Gruppen nach ihrem Rechtsstatus unterschieden wurden:

1. „Bundeseinrichtungen mit FuE-Aufgaben“: Darin sind alle nicht rechtsfähigen Anstalten öffentlichen Rechts zusammengefasst, bei denen die Ministerien Rechts- und Fachaufsicht

4 Daraus erklärt sich, dass es sich um eine inhaltlich und strukturell außerordentlich heterogene Gruppe von Einrichtungen handelt (vgl. Barlösius 2008).

5 Gründe für die verringerte Anzahl sind u.a. der Wechsel des Deutschen Archäologischen Instituts (DAI) zum Auswärtigen Amt sowie die Entscheidung der Bundesregierung, das Institut für Sozialarbeit und Sozialpädagogik e.V. (ISS) fortan nicht mehr zu den Einrichtungen mit Ressortforschungsaufgaben zu rechnen.

6 Die Neuordnung hat das BMELV in dem „Konzept für eine zukunftsfähige Ressortforschung im Geschäftsbereich der BMELV“ dargelegt. Hauptsächlich wurden Einrichtungen zusammengelegt, daneben wurde ein Institut neu gegründet: das Deutsche Biomasseforschungszentrum (eine GmbH).

haben. Diese Institute unterliegen „einer hierarchischen Organisationsstruktur und der Fachaufsicht des zuständigen Ministeriums“ (BMBF 2007b: 4).

2. „Kontinuierliche Zusammenarbeit mit FuE-Einrichtungen“: Diese Kategorie umfasst alle Einrichtungen mit anderen Rechtsformen (z.B. gGmbH, Vereine, Stiftung). Diese Einrichtungen – so konstatiert das Konzept der Bundesregierung – „entsprechen häufig eher dem Typ außeruniversitärer Forschungseinrichtungen mit entsprechender Rechtsform“ (BMBF 2007b: 4).

Die Kategorisierung nach der Rechtsform steht in einem deutlichen Spannungsverhältnis zu dem, was der WR über die Bedeutsamkeit der Rechtsformen festgestellt hat: „Alle vorhandenen Rechtsformen in der Ressortforschung erscheinen demnach prinzipiell geeignet, dem wissenschaftlichen Personal hinreichend Autonomie für ihre FuE-Arbeiten im Rahmen des Ressortforschungsauftrags zu gewähren.“ (WR 2007: 90) Empirisch ist zwischen dem Rechtsstatus und den Leistungen der Einrichtungen kein Zusammenhang erkennbar, weder bezüglich des *Leistungsspektrums*, z.B. Forschungsorientierung oder Ausmaß hoheitlicher Aufgaben, noch bezüglich der *Qualität* der Leistungen. Auch im Hinblick auf die ministerielle Durchgriffstiefe – Rechts- und/oder Fachaufsicht – zeigt sich in der Praxis mehrheitlich keine große Differenz. Nur wenige Ministerien nutzen den direkten Zugriff umfassend zur Ausrichtung und Kontrolle der Einrichtungen. Bei vielen Instituten der zweiten Gruppe ist zu beobachten, dass die Ministerien durch die Mittelzuweisung einen Zugriff auf die Forschungsagenda haben und auf diese Weise die Institute inhaltlich ausrichten (vgl. Barlösius 2010).

Die von den Ministerien neu eingeführte Kategorisierung nach dem formalen Kriterium Rechtsstatus verdeutlicht, dass eine Ordnungsweise geschaffen wurde, die nach dem Grad der Zugehörigkeit zum staatlichen Feld unterscheidet. Gar nicht herangezogen wurden Kriterien, die eine Zugehörigkeit zum wissenschaftlichen Feld bestätigen könnten, wie beispielsweise der Umfang eigener Forschung. Stattdessen wurden Weisungsrechte, das heißt die ministerielle Durchsetzungsfähigkeit, zum entscheidenden Kriterium. Die Ministerien verorten folglich die Ressortforschungseinrichtungen im staatlichen Feld.

1.2 Benennung der Einrichtungen

Die Macht, Benennungen durchzusetzen – darauf hat Pierre Bourdieu aufmerksam gemacht –, stellt eine zentrale Machtquelle dar. Diese Macht – Benennungsmacht – realisiert sich in Kategorisierungs- und Klassifizierungsarbeit, die häufig in Auseinandersetzungen darüber mündet, wie die „Dinge“ richtig zu titulieren seien (Bourdieu 1985). Dafür bietet die Evaluation der Ressortforschungseinrichtungen anschauliches Material. Über die Titulierung der Einrichtungen ist in der ersten Phase des Evaluationsprozesses, als die Einrichtungen des BMELV betrachtet wurden, nicht gesprochen worden. Ihre Benennungen waren nicht „umkämpft“. In den Empfehlungen ebenso wie in den Dokumenten wurden sie „Ressortforschungseinrichtungen“ genannt (WR). Erst danach wurde die Titulierung überhaupt ein Gegenstand der Auseinandersetzung. Der WR hat – abgesehen von seinen Veröffentlichungen, die die offiziellen Dokumente der Evaluation repräsentieren – stets die Begriffe „Ressortforschung“ oder „Ressortforschungseinrichtungen“ verwendet, ersteren jedoch als Synonym oder Kurzform des letzteren und nicht, um einen eigenen Forschungstypus zu kennzeichnen. In den Namen der WR-Arbeitsgruppen und WR-Ausschüsse, im Arbeitsprogramm und Organigramm tauchten diese Begriffe auf.

Im Bundesbericht Forschung war der Abschnitt jeweils mit „Bundeseinrichtungen mit FuE-Aufgaben“ überschrieben. Im Evaluationsauftrag an den WR, verfasst vom Haushaltsausschuss des Deutschen Bundestages, wurde dagegen die Bezeichnung „Ressortforschungseinrichtungen“ ohne Zusatz gebraucht. Die Einrichtungen selbst bezeichnen sich durchweg als Ressortforschungseinrichtungen. So heißt beispielsweise der 2005 gegründete Zusammenschluss der

Einrichtungen „Arbeitsgemeinschaft der Ressortforschungseinrichtungen". Die Empfehlungen des WR von 2007 wiederum verwenden jedoch den Begriff „Bundeseinrichtungen mit FuE–Aufgaben" und die aus dem Jahr 2010 die Bezeichnung „Einrichtungen mit Ressortforschungsaufgaben des Bundes". Der Gebrauch verschiedener Benennungen geschah nicht aus Unachtsamkeit. Im Gegenteil, er resultiert aus Verhandlungen darüber, wie diese Institute in ihrer Gesamtheit zu titulieren seien, bei denen sich der Bund, vertreten durch das BMBF, durchgesetzt hat. Man könnte versucht sein, die Umbenennungen als bedeutungslos zu bewerten. Dies wäre aber voreilig. Erstens dokumentiert sich darin, dass der Bund, sprich die Ministerien, die Macht der Benennung für sich beansprucht. Zweitens zeigt sich bei genauerer Betrachtung der drei Bezeichnungen – Ressortforschungseinrichtungen, Bundeseinrichtungen mit FuE-Aufgaben und Einrichtungen mit Ressortforschungsaufgaben des Bundes –, dass die Bedeutung der Forschung immer weiter relativiert wurde. Bei der ersten Titulierung ist die Forschung das Charakteristikum dieser Einrichtungen und distanziert sie von anderen Bundeseinrichtungen, bei der zweiten ist Forschung zu einer unter anderen Aufgaben geworden, und die dritte bindet Forschungstätigkeiten unmittelbar an Ressortzuständigkeiten, was eine deutliche Einengung der Forschung im Vergleich mit der ersten, aber auch mit der zweiten Bezeichnung impliziert.[7] Auch die Umbenennungen weisen in Richtung einer stärkeren Bindung der Einrichtungen an das staatliche Feld.

1.3 Klassifikation der Leistungen

Bei der Klassifikation der Leistungen belegt „eigene Forschung" die Zugehörigkeit zum wissenschaftlichen Feld, dagegen sind „hoheitliche Tätigkeiten" als Ausdruck der Zugehörigkeit zum staatlichen Feld zu werten. Zu fragen ist deshalb, ob und in welchem Umfang die Leistungen der Ressortforschungseinrichtungen als Erfüllung hoheitlicher Aufgaben eingestuft werden und ob sich dies während der acht Jahre der Evaluation verändert hat. Wird die Hoheitlichkeit als primäres Merkmal aufgefasst, unabhängig davon, ob es sich um Forschung, Beratung oder Prüfaufgaben handelt, dann wird darüber eine deutliche Differenz zum wissenschaftlichen Feld und gleichzeitig eine enge Bindung an das staatliche Feld behauptet.

Über die ministerielle Klassifikation gibt der Bundesbericht Forschung eine erste grobe Auskunft. In den Berichten von 2004 und 2006 findet sich dazu: „Die Bundeseinrichtungen mit Forschungsaufgaben [...] nehmen ihre hoheitliche Tätigkeit im Kontext der Aufgaben des Bundesministeriums […] wahr" (BMBF 2004: 118, 2006: 101). Zwei Jahre später wurde die Klassifikation verändert: „Die Bundeseinrichtungen mit Forschungs- und Entwicklungsaufgaben nehmen ihre Tätigkeit im Kontext der Aufgaben des jeweiligen Bundesministeriums […] wahr" (BMBF 2008: 146, 2010: 60). Während noch in den Bundesberichten von 2004 und 2006 die Bundesregierung sämtliche Aufgaben der Ressortforschungseinrichtungen als hoheitlich klassifiziert und damit fest im staatlichen Feld verankert hat, werden seit 2008 nicht mehr alle Tätigkeiten per se als hoheitlich qualifiziert: Der Begriff kommt nicht einmal mehr vor. Auch in den 2007 von der Bundesregierung herausgegebenen „Zehn Leitlinien einer modernen Ressortforschung" (BMBF 2007a) wie auch in dem „Konzept einer modernen Ressortforschung" (BMBF 2007b) sucht man den Begriff „hoheitlich" vergeblich, ebenso in dem „Konzept für eine zukunftsfähige Ressortforschung BMELV" (BMELV 2004). Das heißt: In den offiziellen von der Bundesregierung verantworteten Texten, Gesetzen und Leitlinien wird hoheitlich nicht mehr als besonderes Kennzeichen der Tätigkeiten der Einrichtungen angeführt.

In seinen Empfehlungen zu den Ressortforschungseinrichtungen des BMELV hat sich der WR mit der Hoheitlichkeit von Tätigkeiten auseinandergesetzt und verdeutlicht, dass der Begriff auf einige wenige Tätigkeiten zu begrenzen sei. Keineswegs dürfe er auf die gesamte Einrichtung bezogen werden. Daraus, dass die Institute auch hoheitliche Aufgaben erfüllen, dürfe keine „spezifische

7 Wie dies mit der zweiten Kategorie „Kontinuierliche Zusammenarbeit mit FuE-Einrichtungen" zu vereinbaren ist, scheint klärungsbedürftig.

Logik“ der Einrichtungen hergeleitet werden. In den WR-Empfehlungen von 2007 findet sich der Begriff „hoheitlich“ nur noch an einer Stelle. Dort wird dargelegt, dass die den Einrichtungen übertragenen staatlich-regulierenden Aufgaben nicht geeignet seien, daraus eine Eigenart zu begründen. „Hoheitliche“ Aufgaben stellen nicht „das eigentliche Proprium der Ressortforschung“ dar (WR 2007: 28). Ansonsten steht der Begriff noch in einer Übersicht, die der WR an die Einrichtungen verschickt hatte, damit diese selbst ihre Leistungen klassifizieren. Im WR-Bericht von 2010 wird der Begriff „hoheitlich“ lediglich noch in der aktualisierten Übersicht der Leistungen verwendet.

Die Ressortforschungseinrichtungen – insbesondere die des BMELV – orientierten sich zunächst an einem Begriff von Hoheitlichkeit, der alle Tätigkeiten umfasste. Eine häufige Selbstbeschreibung war: „Außerhalb der hoheitlichen Aufgaben wird nicht geforscht, insofern erübrigt sich die Frage nach dem Verhältnis von hoheitlichen Aufgaben und Forschung“ (Zitat aus der Begehung vor Ort). Sehr deutlich lässt sich der Wandel der Klassifikation ihrer Tätigkeiten aus den Antworten der Ressortforschungseinrichtungen auf die Bitte des WR ablesen, für die verschiedenen Leistungsbereiche jeweils selbst anzugeben, wie viel Prozent ihrer Tätigkeit diese beanspruchen. Zu diesen Leistungsbereichen zählen eigene Forschung und Entwicklung, Informationsbeschaffung und Politikberatung, hoheitliche Aufgaben, Bereitstellung von Dienstleistungen und Ausbildung. Die Bitte hat der WR 2004 und 2008 an die Institute gerichtet (WR 2007, 2010). Je nach Verteilung der Prozentwerte auf die verschiedenen Leistungsbereiche beschreiben sie sich in ihrem Selbstverständnis eher als Forschungsinstitut, Beratungseinrichtung oder hoheitliche Behörde.[8] Es handelt sich um selbst geschätzte Prozentzahlen. 28 von den 45 Einrichtungen gaben starke Veränderungen an, um mindestens 10 Prozent in geringstenfalls einem Leistungsbereich, meist in zwei Bereichen. Die wenigsten und geringsten Veränderungen teilten die Einrichtungen mit, die bereits bei der ersten Umfrage hohe Werte im Bereich Forschung angegeben hatten, und jene Institute, die schon früher einmal vom WR evaluiert worden waren.

Die größten und die meisten Erhöhungen wurden bezüglich des Forschungsanteils gemeldet. Elf Institute gaben Anstiege um mehr als 10 Prozent an. Nicht wenige bekundeten Steigerungen des Anteils an eigener Forschung um 100 bis 200 Prozent, ein Institut erhöhte seinen Forschungsanteil von 0 auf 34 Prozent. Am stärksten gesunken sind nach Angabe der befragten Einrichtungen die Anteile an hoheitlichen Aufgaben. Bei der Informationsbeschaffung und der Politikberatung sind mehr Anstiege als Abstiege zu verzeichnen, bei der Bereitstellung von Dienstleistungen verhält es sich genau umgekehrt. Die insgesamt gravierenden Veränderungen bei den Selbstklassifizierungen der Leistungen sind nur sehr begrenzt auf strukturelle Veränderungen zurückzuführen, etwa Aufbau oder Schließung von Forschungsabteilungen, Übernahme oder Abgabe hoheitlicher Aufgaben etc. In der überwiegenden Mehrzahl der Fälle haben die Einrichtungen ihre Leistungen lediglich anders klassifiziert.

Viele Einrichtungen, die 2004 niedrige Prozentzahlen für den Forschungsanteil meldeten, haben diese 2008 höher gesetzt. Ein Motiv, den Forschungsanteil sehr gering anzusetzen, war vermutlich, dass solche Einrichtungen die Evaluation durch den WR, der primär die Qualität der Forschung bewertete, als für sich unangemessen zurückweisen wollten. Manchen dieser Einrichtungen scheint jedoch im Laufe der Evaluation deutlich geworden zu sein, dass vom WR bescheinigte gute bis sehr gute Forschungsleistungen einen wissenschaftlichen Reputationsgewinn verschaffen, der sich in eine gesteigerte Anerkennung umsetzen lässt. So ließ sich für einige Institute eine Höherschätzung sowohl durch das eigene Ministerium, innerhalb der Gruppe der Ressortforschungseinrichtungen als auch durch das wissenschaftliche Feld beobachten. Weiterhin scheint sich bei diversen Einrichtungen die Überzeugung durchgesetzt zu haben, dass diese Anerkennung ihren Bestand und ihre Ausstattung augenscheinlich besser garantiert als der Verweis darauf, dass sie hoheitlichen

8 Bei Einrichtungen, die angeben, dass sie keine oder zu weniger als 10 Prozent Forschung betreiben – einerlei, ob diese Angabe dem realen Anteil entspricht oder unterschätzt wird –, ist zu fragen, ob es sich um eine Ressortforschungseinrichtung handelt oder nicht vielmehr um ein Bundesamt.

Aufgaben nachkommen. Nicht zuletzt lässt sich daraus auch das Anrecht eines gewissen Maßes an Unabhängigkeit vom Ministerium herleiten.

Aus der Erhöhung der Prozentangaben für die eigene Forschung könnte man schließen, dass die Einrichtungen näher an das wissenschaftliche Feld herangeführt werden sollen. Diese Veränderungen bei der Klassifikation der Leistungen stehen scheinbar im Widerspruch zu der erwähnten neuen Unterscheidung nach der Rechtsform und den Umbenennungen der Gesamtheit der Einrichtungen. Beide Veränderungen fassen die Ressortforschungseinrichtungen eindeutiger als vor Beginn des Evaluationsprozesses als Teil des staatlichen Felds auf. Allerdings wird bei den Leistungen die Zugehörigkeit nicht mehr mittels der Hoheitlichkeit der Tätigkeiten, sondern stärker durch ihr Gerichtetsein auf staatliche (ministerielle) Aufgaben begründet.[9]

Diese Ebene wurde ausführlicher dargestellt, als dies für die beiden weiteren vorgesehen ist. Der Grund dafür ist, dass sich aus der Positionierung schon zu großen Teilen herleitet, ob sich die interne Strukturierung eher an staatlichen Behörden oder an wissenschaftlichen Instituten orientiert. In ähnlicher Weise gilt dies für die dritte Ebene, die Ausformulierung eines eigenen Nomos.[10]

2 Interne Strukturierung sowie Selbstkontrolle und Selbstausrichtung

Die von den Evaluatoren vorgefundene Ausgestaltung der internen Strukturierung ebenso wie die ministeriell garantierten und von den Einrichtungen ergriffenen Spielräume zur Selbstkontrolle und -ausrichtung unterscheiden sich erheblich. Die Ressortzugehörigkeit stellt sich dabei als besonders strukturierend heraus. Einige Ministerien unterstützen ihre Institute dabei, gemeinsame Berufungen mit Universitäten durchzuführen, andere Ministerien halten dies für nicht verfassungsgemäß oder für haushaltsrechtlich problematisch. Es gibt Ministerien, die Befristungen grundsätzlich ablehnen, andere erlauben es, dass ihre Einrichtungen das Wissenschaftszeitvertragsgesetz nutzen. Fragt man, ob im Zusammenhang mit den Evaluationen Veränderungen auf dieser Ebene stattgefunden haben, so ergibt sich ein widersprüchliches Bild.[11] Anhand von drei Punkten soll dies dargestellt werden:[12]

1. Der WR ist dafür bekannt, dass er ein bestimmtes Repertoire von Instrumenten, Institutionen und Organisationsformen für geeignet hält, wissenschaftsfördernde Strukturen zu schaffen. Dazu gehören beispielsweise befristet beschäftigte Wissenschaftlerinnen und Wissenschaftler, gemeinsame Berufungen außeruniversitärer Einrichtungen mit Hochschulen, die Einrichtung von wissenschaftlichen Beiräten, Globalhaushalte etc. Nach der Veröffentlichung der ersten WR-Bewertungsberichte war zu beobachten, dass einige Einrichtungen und Ministerien schon vor der Evaluation die Einrichtung eines wissenschaftlichen Beirats angekündigt oder diesen bereits installiert hatten.[13] Die Einführung der anderen Instrumente, Institutionen und Organisationsformen wurde dagegen wesentlich seltener in Aussicht gestellt.
2. Mit den „Zehn Leitlinien einer modernen Ressortforschung“ (BMBF 2007a) und dem „Konzept einer modernen Ressortforschung“ (BMBF 2007b) reagierten die Ministerien

9 Ob ein und welcher Wandel des Staatsverständnisses sich dahinter verbirgt, ist eine interessante politikwissenschaftliche Frage.

10 Deswegen werden für die beiden Ebenen die vorne benannten Punkte nicht systematisch abgehandelt.

11 Endgültige Schlüsse zu ziehen wäre auf dieser Ebene vorschnell, weil Veränderungen der Administration und rechtlicher Regelungen längere Zeiträume benötigen.

12 Eine geeignete Quelle, um nachzuprüfen, ob und welche strukturellen Veränderungen vorgenommen wurden, sind die sogenannten Nachverfolgungen. Hier berichten die Einrichtungen, welche WR-Empfehlungen umgesetzt, welche abgelehnt und welche für die Zukunft geplant sind. Diese habe ich jedoch nicht systematisch ausgewertet.

13 Die Einrichtung eines wissenschaftlichen Beirats verlangt vergleichsweise geringe institutionelle Veränderungen.

unmittelbar auf die WR-Empfehlungen von 2007. Das Konzept stellt wissenschaftsspezifische Kategorien, Institutionen und Strukturen vor, verwendet wissenschaftstypische Begriffe und Argumente und zitiert wissenschaftspolitische Überzeugungen, allerdings mehrheitlich nur als Soll-Vorstellungen, ohne Verbindlichkeit für die Ministerien und die Einrichtungen. Außerdem werden die Empfehlungen des WR in dem Konzept nur auf jene Einrichtungen bezogen, „die in hohem Maße eigene Forschung und Entwicklung betreiben"; in diesen „muss wissenschaftsspezifischen Belangen in besonderem Maße Rechnung getragen werden" (BMBF 2007b: 3). Insgesamt bestätigt das Konzept die alleinige Zuständigkeit der Ministerien für die institutionelle und organisatorische Ausrichtung ihrer Einrichtungen und setzt sich nicht für ein übergeordnetes wissenschaftspolitisches Gestaltungsrecht ein. Demgemäß wird keine Perspektive aufgezeigt, die Ressortforschungseinrichtungen zu einer eigenen Säule der Wissenschaftsorganisationen weiterzuentwickeln.

3. Unter dem Aspekt der Schaffung von Institutionen der Selbstausrichtung ist die Gründung der „Arbeitsgemeinschaft der Ressortforschungseinrichtungen" im Jahr 2005 als Organ der Selbstorganisation der Ressortforschungseinrichtungen bedeutsam. Ein wichtiger Anstoß dafür war die Evaluation durch den WR. Die Gründung kann als ein erster Schritt, eine eigene Wissenschaftsorganisation zu etablieren, angesehen werden. Der Arbeitsgemeinschaft gehören gegenwärtig 39 der 46 Einrichtungen an.[14] Sie sieht die Ressortforschungseinrichtungen als Teil des wissenschaftlichen Feldes und setzt sich dafür ein, dass diese nach den „üblichen akademischen Kriterien der Qualitätsbewertung" beurteilt werden.[15]

Zusammenfassend lässt sich festhalten, dass die Evaluationen zu einer größeren Offenheit sowohl der Ministerien wie auch der Ressortforschungseinrichtungen gegenüber den wissenschaftstypischen Instrumenten, Institutionen und Organisationsformen geführt haben. Dies gilt insbesondere für Institute, die sich nah am wissenschaftlichen Feld positionieren und auch von den zuständigen Ministerien dort gesehen werden.

3 Wissenschaftlicher Nomos

Ob die Leistungen der Ressortforschungseinrichtungen einem originären Nomos entsprechen, aus dem sich eine spezifische Logik und damit ein eigenes Feld begründen lässt, ist ein Aspekt, der den gesamten Prozess der Evaluation begleitet hat. Diskutiert wurde vor allem entlang der Frage, ob Ressortforschung einen eigenen Forschungstypus repräsentiert. Wird dies zuerkannt, dann rechtfertigt auch die Forschung, und nicht nur das Erfordernis eines privilegierten ministeriellen Zugriffs auf Wissenschaft, die Errichtung spezieller Institute (vgl. Barlösius 2010).

Für den WR resultierten aus der Frage nach einem eigenen Nomos zwei Herausforderungen: Erstens hat er an seinen im wissenschaftlichen Feld etablierten und wissenschaftspolitisch anerkannten Kriterien für „gute Wissenschaft" festzuhalten. Zweitens muss er zu einer Betrachtungsweise der Ressortforschungseinrichtungen gelangen, die von den verschiedenen Ministerien, den Instituten selbst, der Politik, aber auch den wissenschaftspolitisch engagierten Akteuren des wissenschaftlichen Feldes anerkannt wird. Der erste Punkt ist wichtig, weil davon die Fähigkeit des WR abhängt, seine Kriterien für das gesamte wissenschaftliche Feld durchzusetzen und darüber seine dominante Position – als oberster Begutachter der Strukturen und Organisation von Wissenschaft – zu behaupten. Der zweite Punkt ist bedeutsam, weil ansonsten die Zustimmung der Bundesregierung zu den WR-Empfehlungen und somit ihre Verabschiedung und Veröffentlichung fraglich ist. Seinem Rang als

14 Bis auf einige wenige forschungsfernere Institute, z.B. das Bundesinstitut für Sportwissenschaft, sind beinahe alle Ressortforschungseinrichtungen Mitglied der Arbeitsgemeinschaft. Eine Ausnahme bilden die drei Ressortforschungseinrichtungen des Bundesministeriums für Umwelt (UBA, BfN und BfS); sie wirken in der Arbeitsgemeinschaft nicht mit.

15 Siehe http://www.ressortforschung.de/de/home/index.htm.

Konsensorgan zwischen Wissenschaft und Politik entspricht es, Empfehlungen zu formulieren, die auf breite Zustimmung treffen.

Die Frage, ob die Forschung in diesen Einrichtungen einen eigenen Forschungstypus repräsentiert, hat der WR folgendermaßen beantwortet: „Die Forschung der Bundesanstalten selbst kann [...] keinen Sonderstatus für sich beanspruchen. Sie ist Teil des Wissenschaftsdiskurses im jeweiligen Fachgebiet und unterliegt den gültigen Anforderungen an die Forschung in den jeweiligen Fachgebieten. Maßstab ist der ‚state of the art'" (WR 2004: 48). Auch die „Langfristaufgaben und Aufgaben der Politikberatung stellen [...] kein Alleinstellungsmerkmal der Ressortforschung dar" (WR 2004: 56). Mit dem Hinweis darauf, dass auch andere wissenschaftliche Institute solche Leistungen erbringen, die als „eigentliches Proprium der Ressortforschung gelten", wies der WR einen durch die Spezifik der Aufgaben begründeten Sonderstatus der Ressortforschung zurück und unterstrich die Nähe zu anderen wissenschaftlichen Instituten. An dieser Position hat der WR konsequent über den gesamten Zeitraum festgehalten. Von dem Grundsatz „Forschung ist Forschung" leitet er her, dass „kein eigener Forschungsbegriff für die Ressortforschung" existiert. Überall, wo geforscht wird, so der Standpunkt des WR, gelten prinzipiell dieselben Kriterien für gute Forschung, weshalb die Ressortforschungseinrichtungen auch keine Sonderstellung für sich beanspruchen können. Entsprechend hat der WR in seinen Berichten (bis auf die WR-Empfehlungen von 2010) den Begriff der Ressortforschung gemieden, um Ambitionen, einen speziellen Forschungstypus zu konstituieren, entgegenzuwirken. Damit hat der WR der Forschung in den Ressortforschungseinrichtungen einen eigenen Nomos abgesprochen und die dort geleistete Forschung dem wissenschaftlichen Feld zugerechnet. Weiterhin hat er die Qualität der anderen Leistungsbereiche, z.B. Politikberatung und Normung, als unmittelbar durch die Forschungsqualität bedingt bestimmt und damit einen Primat der Forschung vor allen anderen Leistungen formuliert. Allerdings gebraucht der WR die unmittelbare Verknüpfung von Forschung und Entwicklung mit anderen Aufgaben innerhalb der Einrichtungen als eine überzeugende Begründung dafür, dass die Ministerien für ihre Forschungs- und Entwicklungszwecke über eigene Institute verfügen (vgl. auch die Stellungnahmen WR 2007 und 2010).

In den „Zehn Leitlinien einer modernen Ressortforschung" (BMBF 2007a) wie auch in dem „Konzept einer modernen Ressortforschung" (BMBF 2007b) hat die Bundesregierung dazu eine Gegenposition verfasst. Sie lautet: „Ihre Fähigkeit, Wissenschaft, Politikberatung und Vollzug miteinander zu verknüpfen und für das Regierungshandeln aufzubereiten, macht die Besonderheit und das Alleinstellungsmerkmal dieser Einrichtungen aus" (BMBF 2007b: 4). „Forschungs- und Entwicklungsaktivitäten des Bundes, die der Vorbereitung, Unterstützung oder Umsetzung politischer Entscheidungen dienen und untrennbar mit der Wahrnehmung öffentlicher Aufgaben verbunden sind, sind als ‚Ressortforschung' definiert" (BMBF 2007b: 3). Die Ressortforschung forscht „problemorientiert", „praxisnah", „interdisziplinär", „transdisziplinär", „generiert Transferwissen", „erbringt Übersetzungsleistungen", stellt „kurzfristig abrufbare wissenschaftliche Kompetenz" bereit, arbeitet an „langfristig angelegten Fragestellungen" und „agiert in diversen Spannungsfeldern, die durch unterschiedliche Rationalitäten der Wissenschaft und der Politik gekennzeichnet" sind (BMBF 2007b: 3). Diese Auffassung unterstellt die Forschungstätigkeiten vollkommen der Ressortzuständigkeit und räumt der nicht gebundenen Forschung, der „freien Forschung", die sich nicht unmittelbar auf politische Entscheidungen und öffentliche Aufgaben bezieht, keinen Platz ein.

Empirisch, das hat die Evaluation gezeigt, existiert zwischen den verschiedenen Einrichtungen eine große Spannweite hinsichtlich der Frage, ob sie „freie Forschung" für die Gewährleistung „guter Leistungen" für erforderlich halten. Es gibt Ministerien und Einrichtungen, die originäre Forschung für entbehrlich halten, weil sie die Ressortforschung als „Transferriemen von der Wissenschaft in die Politik" betrachten, worunter sie vor allem die „Übersetzung" von Forschung in Politik und die Einbettung der Forschungsergebnisse in gesellschaftliche Kontexte verstehen.

Aus ihrer Perspektive genügt es, wenn die Einrichtungen mit dem Stand der Wissenschaft vertraut sind. Andere halten eigene, selbst bestimmte Forschung für unverzichtbar, weil nur so verlässliche Ergebnisse für die Politikberatung erbracht werden könnten und die eigene Forschungsreputation für die politische Durchsetzungsfähigkeit, insbesondere international, essentiell sei. Die Sicherung der wissenschaftlichen Qualität der Einrichtungen geschieht aus ihrer Sicht über hochrangige Publikationen, Peer Reviews, Evaluationen, den internationalen Ruf etc. Das „Mithalten-Können der Ressortforschung mit dem Rest des Wissenschaftssystems" erachten sie für wichtig.[16] Zwischen diesen beiden Positionen gibt es viele Abstufungen.

Beide Positionen rekurrieren nicht auf einen eigenen Nomos der Ressortforschungseinrichtungen. Vielmehr entsprechen ihre Sichtweisen entweder dem ministeriell festgelegten Anforderungsprofil oder sie orientieren sich am Nomos der Wissenschaft. Die in dem Konzept der Bundesregierung entworfene Definition der Ressortforschung, die einen eigenen Forschungstypus behauptet, wird dagegen kaum vertreten. Ein Grund dafür könnte sein, dass diese aus einem Sammelsurium von Merkmalen besteht. Weder die Einrichtungen noch die Ministerien verwenden diese, wenn sie schildern, was das Besondere ihrer Leistungen ist.

4 Ein kurzes Resümee

Die Ausgangsfrage war, ob die Evaluationen durch den WR zu einer einheitlicheren Auffassung und Beschreibung der Ressortforschungseinrichtungen beigetragen haben. Prüft man, um diese Frage zu beantworten, ob sich im Laufe der acht Jahre der Evaluationen die Sichtweisen und Beurteilungskriterien der drei Akteure der Evaluation – Ressortforschungseinrichtungen, Ministerien und WR – einander angeglichen oder sich weiter auseinanderentwickelt haben, so ergibt sich ein widersprüchliches Bild. Vor allem aber wird deutlich, dass diese Frage zu voraussetzungsvoll formuliert ist. Sie geht davon aus, dass die drei Akteure bereits vor der Evaluation über eine für sie stimmige Beschreibung und Auffassung der Ressortforschungseinrichtungen verfügten. Dies war aber nur bedingt der Fall. So scheint die Evaluation die Ressortforschungseinrichtungen und die Ministerien ermahnt zu haben, eigene Sichtweisen auszuarbeiten und für Dritte überzeugend abzufassen. Die „Leitlinien" sowie das „Konzept" der Bundesregierung und die Selbstdarstellung der „Arbeitsgemeinschaft der Ressortforschungseinrichtungen" sind dafür anschauliche Beispiele. Teilweise mussten dafür Begriffe definiert, Stellungnahmen entwickelt und Abgrenzungen bestimmt werden, insbesondere gegenüber der sprachlichen Dominanz des WR, der reichhaltige Erfahrungen mit Evaluationsprozessen besitzt. Auch der WR hatte die Aufgabe, eine in sich stimmige Beschreibung und Auffassung zu entwickeln, wobei er, und das unterscheidet ihn von den beiden anderen Akteuren, auf eine kohärente Abstimmung mit seinen Empfehlungen zum gesamten wissenschaftlichen Feld zu achten hatte.

Weiterhin, das ist hier zu kurz gekommen, war die Evaluation ein Anlass für die Ministerien und die Ressortforschungseinrichtungen, sich gegenseitig darüber zu informieren, wie sie sich selbst verstehen, welche Art von Forschung sie betreiben, was für sie gute Leistungen sind, ob sie sich am wissenschaftlichen Feld orientieren oder sich als Wissenschaftsbehörde sehen, die zum staatlichen Feld gehört. Durch diesen Austausch wurde die große Heterogenität der Einrichtungen deutlich. Da der WR nicht nur die einzelnen Einrichtungen evaluiert, sondern auch übergreifende Stellungnahmen formuliert hat, stand er vor der Aufgabe, Empfehlungen zu verfassen, die mehr oder weniger auf alle Einrichtungen zutreffen bzw. auf sie angewendet werden können. Diese Aufgabe setzte eine vereinheitlichende Beschreibung der Ressortforschungseinrichtungen durch den WR voraus. Die Leitlinien und das Konzept der Bundesregierung sowie die Selbstbeschreibung durch die „Arbeitsgemeinschaft der Ressortforschungseinrichtungen" haben einen ähnlichen Anspruch verfolgt. Insgesamt hat der Evaluationsprozess alle drei beteiligten Akteure dazu gedrängt, aus ihrer

16 In diesem Absatz beziehe ich mich auf Anhörungen der Ministerien durch den WR.

jeweiligen Perspektive eine in sich stimmige Beschreibung der Ressortforschungseinrichtungen zu verfertigen. Eine gemeinsame Beschreibung und Auffassung der Einrichtungen ist dabei nicht entstanden. Dies war auch nicht zu erwarten. Bei Einrichtungen, deren Feldzugehörigkeit sowohl intern als auch gegenüber anderen Feldern umkämpft ist, ist selbstverständlich auch die Beschreibung umstritten. Eine verbindliche Beschreibung kann nur autorisieren, wer die Benennungsmacht besitzt, und in diesem Fall rang das wissenschaftliche mit dem staatlichen Feld um das Vorrecht der Titulierung, Kategorisierung und Klassifikation. Ganz unabhängig davon ist die größte Veränderung, dass die Ressortforschungseinrichtungen nicht länger eine „terra incognita" sind: Sie wurden und haben sich entdeckt. Die Entdeckten, die Hüter der wissenschaftspolitisch unentdeckten Flecken genauso wie die Entdecker haben sie vermessen und kartografiert. Sie haben dazu unterschiedliche Maßstäbe verwendet und verschiedene topographische Eigenheiten geortet.

Literatur

Barlösius, Eva, 2006: Wissenschaft evaluiert – praktische Beobachtungen und theoretische Betrachtungen, in: *Flick, Uwe (Hg.):* Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek bei Hamburg: rororo, 385-405.

Barlösius, Eva, 2007: Urteilsgewissheit und wissenschaftliches Kapital, in: Wissenschaft unter Beobachtung. Sonderheft des Leviathan 24/2007, 248-264.

Barlösius, Eva, 2008: Zwischen Wissenschaft und Staat? Die Verortung der Ressortforschung. WZB-Discussion Paper P 2008-101. Berlin: WZB.

Barlösius, Eva, 2009: „Forschen mit Gespür für politische Umsetzung" – Position, interne Strukturierung und Nomos der Ressortforschung. dms – der moderne Staat – Zeitschrift für Public Policy, Recht und Management, Heft 2/2009, 347-366.

Barlösius, Eva, 2010: Stichwort: Ressortforschung, in: *Hornbostel, Stefan / Knie, Andreas / Simon, Dagmar (Hg.):* Handwörterbuch Wissenschaftspolitik. Wiesbaden: VS Verlag für Sozialwissenschaften, 377-389.

BMBF, 2004: Bericht des Bundesministeriums für Bildung und Forschung zur Evaluierung der Ressortforschung, Anlage zur BMF-Vorlage Nr. 99/04. Bonn/Berlin.

BMBF, 2004: Bundesbericht Forschung. Bonn/Berlin.

BMBF, 2006: Bundesbericht Forschung. Bonn/Berlin.

BMBF (Hg.), 2007a: Zehn Leitlinien einer modernen Ressortforschung. Bonn/Berlin.

BMBF, 2007b: Konzept einer modernen Ressortforschung. Bonn/Berlin.

BMBF, 2008: Bundesbericht Forschung. Bonn/Berlin.

BMBF, 2010: Bundesbericht Forschung. Bonn/Berlin.

BMELV, 2006: Konzept für eine zukunftsfähige Ressortforschung im Geschäftsbereich des BMELV, 27.09.2006. Bonn.

Bourdieu, Pierre, 1985: Sozialer Raum und „Klassen". Frankfurt/M.: Suhrkamp.

Bourdieu, Pierre, 1997: Les usages sociaux de la science. Paris: INRA.

Chevassus-au-Louis, Bernard, 2007: L'analyse des riques. L'expert, le décideur et le citoyen. Paris: INRA.

Dejours, Christophe, 2003: L'évaluation du travail à l'épreuve du réel. Critique des fondements de l'évaluation. Paris: INRA.

Döhler, Marian, 2007: Die politische Steuerung der Verwaltung. Eine empirische Studie über politisch-administrative Interaktionen auf der Bundesebene. Baden-Baden: Nomos.

Hohn, Hans-Willy / Schimank, Uwe, 1990: Konflikte und Gleichgewichte im Forschungssystem. Akteurskonstellationen und Entwicklungspfade in der staatlich finanzierten außeruniversitären Forschung. Frankfurt/M.: Campus.

Krauss, Gerhard (1996): Forschung im unitarischen Staat. Abhängigkeit und Autonomie der staatlich finanzierten Forschung in Frankreich. Frankfurt/M.: Campus.

Künast, Renate, 2007: „Die Kuh umzingeln". Gespräch mit Wolfert von Rahden und Christoph Mielzarek. Gegenworte, Heft 18, Herbst 2007, 53-56.

Lieske, Jürgen, 2000: Forschung als Geschäft. Die Entwicklung von Auftragsforschung in den USA und Deutschland. Frankfurt/M.: Campus.

Lundgreen, Peter / Horn, Bernd / Krohn, Wolfgang, 1986: Staatliche Forschung in Deutschland 1870-1980. Frankfurt/M.: Campus.

Roqueplo, Philippe, 1997: Entre savoir et decision, l'expertise scientiftique. Paris: INRA.

Schimank, Uwe, 2005: Zukunft der Ressortforschung, Vortrag am 24.02.2005 in Bonn.

Stone, Diane / Denham, Andrew / Garnett, Mark (Eds.), 1998: Think Tanks across Nations. Manchester: Manchester University Press.

Wollmann, Helmut, 2001: Policy Knowledge: Contract Research, in: International Encyclopedia of the Social & Behavorial Sciences. London: Elesevier, 11574-11578.

WR, 2004: Empfehlungen zur Entwicklung der Rahmenbedingungen der Forschung in Ressortforschungseinrichtungen. Am Beispiel der Forschungsanstalten in der Zuständigkeit des BMVEL. Köln.

WR, 2007: Empfehlungen zur Rolle und künftigen Entwicklung der Bundeseinrichtungen mit FuE-Aufgaben. Köln.

WR, 2010: Empfehlungen zur Profilierung der Einrichtungen mit Ressortforschungsaufgaben des Bundes. Köln.

Marc Torka

Institutioneller gleich handlungspraktischer Wandel? Das Beispiel von Begutachtungspraktiken bei der Evaluation wissenschaftlicher Einrichtungen

Verfahren der institutionellen Evaluation von ganzen wissenschaftlichen Einrichtungen und Forschungsfeldern erzeugen eine neue Handlungssituation für Gutachter/innen: Die Gegenstände, der Prozess und die soziale Situation der Urteilsfindung verändern sich im Vergleich zu anderen Formen der Begutachtung. Wandeln sich in dieser neuartigen Handlungssituation aber auch die Begutachtungspraktiken und Wertorientierungen von Gutachterinnen und Gutachtern? Oder werden diese nur auf neue Handlungssituationen übertragen? Um der Frage nachzugehen, inwieweit institutionelle Veränderungen der Wissenschaft auch zu neuen Handlungsweisen führen, werden Praktiken des Begutachtens im Kontext von zwei Evaluationstypen analysiert.

1 Das Problem, den institutionellen Wandel von Wissenschaft im Handeln nachzuweisen

Die institutionelle Umwelt der Wissenschaft unterliegt gegenwärtig einem umfangreichen Neuordnungsprozess. Vielfältige wissenschaftspolitische Initiativen zielen auf die Reorganisation von Studiengängen, Karrieren, Förderungsinstrumenten, Organisationsstrukturen und den hier diskutierten Bewertungssystemen der Wissenschaft ab. Hierfür sind die seit den 1980er Jahren vermehrt auftretenden Evaluationen von ganzen Wissenschaftsorganisationen und Wissenschaftsfeldern ein besonders intensiv diskutiertes Beispiel. Die vergangenen Leistungen und zukünftigen Potentiale von wissenschaftlichen Einrichtungen werden damit turnusmäßig überprüft, um über die Weiterfinanzierung und Fortsetzung von Forschungsprogrammen zu entscheiden. Für die Wissenschaftsforschung deuten solche von außen auferlegten Wissenschaftsevaluationen auf einen generellen Vertrauensverlust in die Selbststeuerungsmechanismen der Wissenschaft hin. Dieser mündet in einem gesteigerten Rechtfertigungsbedarf, dem mit Evaluationen begegnet wird (Weingart 2005, Schimank 2005). Bereits die Existenz von Wissenschaftsevaluationen liefert demnach einen Hinweis auf veränderte Autoritätsbeziehungen innerhalb der Wissenschaft und zwischen Wissenschaft und Politik (Whitley/Gläser/Engwall 2010).

Der vorherrschende wissenschaftssoziologische und wissenschaftspolitische Diskurs geht also von einer „New Balance of Power" aus und fragt vor allem danach, welche Folgen Evaluationen für die wissenschaftliche Praxis haben. Das Verhältnis zwischen institutioneller Gestalt und Praxis der Wissenschaft steht somit erneut zur Debatte.[1] Die dominante Vorstellung ist dabei, dass die Institutionalisierung von Evaluationsverfahren mehr oder minder ungebrochen zu einer Neuordnung wissenschaftlicher Wertorientierungen, Handlungs- und Bewertungsweisen führt. Evaluationen setzen formale Rahmenbedingungen und verwenden Entscheidungskriterien, an die sich das wissenschaftliche Handeln dann anpasst. So ziehen Ben Martin und Richard Whitley (2010) beispielsweise den Schluss, dass Wissenschaftsevaluationen den Wettbewerb um Publikationschancen und Forschungsmittel sowie die Herausbildung von disziplinären Eliten beförderten und deshalb generell mit einem „decline in collegiality" zu rechnen sei. Auch verkürzte Publikationsintervalle, die durch einen generellen Publikationsdruck, befristete Drittmittelforschung oder turnusmäßige Evaluationen erzeugt werden, brächten folglich „short term"-, „incremental"- und „mainstream"-

1 Das ist die klassische Konfliktlinie seit der Entstehung der Wissenschaftssoziologie. Die von Latour und Woolgar bzw. Knorr-Cetina begründeten mikrosoziologischen Laborstudien kritisierten seit den 1970er Jahren den zuvor dominanten institutionalistischen Ansatz von Merton, der die Norm- und Wertebasis, aber weniger das konkrete wissenschaftliche Handeln in den Blick nahm. Die „Erneuerung der institutionalistischen Wissenschaftssoziologie" (Schimank 1995) zielt hingegen eher auf rechtliche und formal organisatorische Regulierungen ab.

Forschungen hervor (ebd.: 70f.). Studien über die Grenzen des Einflusses von Evaluationen auf den Kern wissenschaftlichen Handelns sind hingegen rar.[2]

So plausibel und wichtig diese institutionalistische Forschungsperspektive ist, so gering ist bislang das empirisch gesicherte Wissen über die handlungspraktischen Folgen von Evaluationen und so groß sind die analytischen Herausforderungen zur Erforschung dieses Zusammenhangs (Gläser et al. 2008). Die Schwierigkeit besteht nämlich darin, „kausale Mechanismen" zu identifizieren und zu isolieren, die die Regeln von sporadischen Evaluationsereignissen in den Alltag wissenschaftlichen Handelns importieren (Gläser/Laudel 2007). Gelingt eine Isolierung von Mechanismen – zum Beispiel, Leistung nach Drittmittelquoten oder dem Impact von Publikationen zu berechnen, die Bereitstellung von Forschungsmöglichkeiten an diesen Kriterien auszurichten und auf diese Weise der Forschungspraxis einen Orientierungsrahmen für ‚werthaltige' Beiträge zu geben – dann bleibt die Frage offen, inwiefern diese Mechanismen ursächlich mit der Etablierung von Wissenschaftsevaluationen zusammenhängen.[3] Wie noch zu zeigen ist, sind Publikations- und Drittmittelraten noch nicht einmal die bestimmenden Kriterien bei der Urteilsfindung, sondern nur eine unter vielen Informationsquellen der Gutachter/innen.[4] Die institutionalistische Perspektive neigt außerdem dazu, die Reaktions- und Umgangsweisen von Wissenschaftler/innen mit neuartigen Evaluationsregimen unterzubelichten, obwohl diese „by no means just passive recipients of such changes in their institutional environment" (Leisyte/Enders/de Boer 2010: 267) sind. Wir werden sehen, dass gerade in der aktiven Auseinandersetzung mit neuen Regulierungsformen tief verankerte Eigenregulierungen von Wissenschaftler/innen zutage treten, die ihr Handeln strukturieren.

Vor dem Hintergrund dieser ungelösten Schwierigkeiten, die ‚Wirkung' gelegentlicher Evaluationsereignisse auf das Alltagshandeln von Wissenschaftler/innen zu analysieren, schlage ich im Folgenden einen bescheideneren Weg ein. Ich rücke eine Handlungssituation ins Zentrum, in der durch Evaluationen gestiftete Erwartungsstrukturen direkt auf Eigenregulierungen von Wissenschaftler/innen treffen: Anhand der Begutachtungsweisen von Wissenschaftlern im Rahmen von institutionellen Evaluationen frage ich, durch welche Regeln das Gutachterhandeln bestimmt ist. Dort ist es möglich, in situ und ohne verschiedene Handlungssituationen vermittelnde Mechanismen der Frage nachzugehen, inwieweit sich das Handeln von wissenschaftlichen Gutachtern dem institutionellen Rahmen, eigenen Kriterien der Angemessenheit oder einer Gemengelage aus beiden fügt. Da ich mit Praktiken der Urteilsfindung nur einen Aspekt wissenschaftlichen Handelns analysiere, können natürlich keine Schlussfolgerungen gezogen werden, ob sich das konkrete Forschungshandeln (z.B. die Problemwahl) an Evaluationskriterien ausrichtet oder nicht.

Anhand von zwei stark kontrastierenden Evaluationstypen gehe ich der Frage nach, ob sich im Kontext dieser neuen Handlungssituationen auch das Gutachterhandeln in substantieller Weise ändert und welcher Art diese Anpassung gegebenenfalls ist. Dafür werde ich zunächst auf die Evaluationstypen, die Datenbasis und die Methodik eingehen (2), dann der Frage nachgehen, inwieweit im Kontext dieser Evaluationen qua Verfahren (3) und entlang der Situationsdeutung von Gutachtern (4) überhaupt eine grundsätzlich neue Handlungssituation entsteht. Schließlich werde ich analysieren, mit welchen Begutachtungspraktiken diese gegebenenfalls neuen Herausforderungen bewältigt werden (5). Zum Schluss ziehe ich ein Fazit bezüglich der Frage, ob die Regeln der Evaluation sich im Handeln der Gutachter/innen niederschlagen und dieses modifizieren (6).

2 Für das Beispiel der Problemwahl vgl. jedoch Leisyte/Enders/de Boer 2010.

3 Stefan Hornbostel hat in seinem Tagungsbeitrag darauf aufmerksam gemacht, dass die Publikationsraten bereits im Zuge der Expansion des Wissenschaftssystems seit den 1960er Jahren rasant zugenommen haben und nicht erst mit der flächendeckenden Etablierung von Evaluationen ab Mitte der 1980er Jahre. Die Projektform als Grundlage der Drittmittelforschung ist ein noch älteres und in allen Wissenschaftsarten vorkommendes Phänomen (Torka 2009).

4 Jochen Gläser und Grit Laudel (2007) haben mit Australien einen besonders extremen Fall von Wissenschaftsevaluationen analysiert. Bis 2008 wurden ausschließlich metrische Daten verwendet. Seit 2010 findet sich aber auch dort ein dem britischen Research Assessment Exercise vergleichbares Peer Review-Verfahren.

2 Datenbasis, Evaluationstypen, Methodik

Bei meinen Ausführungen stütze ich mich auf reichhaltiges Datenmaterial, das im Kontext eines Forschungsprojekts über Urteilsprozesse in drei Evaluationsverfahren erhoben wurde.[5] Es handelt sich um das niederländische *Standard Evaluation Protocol*, das Evaluationsverfahren der deutschen Leibniz-Gemeinschaft und das britische *Research Assessment Exercise* von 2008. Mit allen diesen Verfahren sollen die Qualitäten wissenschaftlicher Organisationseinheiten überprüft und weiterentwickelt sowie Finanzierungsentscheidungen vorbereitet werden. Dennoch verfolgen sie die gleichen Ziele in unterschiedlicher Weise.

Es lassen sich zwei Grundtypen unterscheiden. Das britische Verfahren evaluiert alle nationalen wissenschaftlichen Einrichtungen zeitgleich, vergleichend und aus aktenkundiger Distanz und erzeugt letztlich ein numerisches Ranking, das zwar der Politik als Verteilungsschlüssel dient, aber kaum eine inhaltliche Rückmeldung an die Institute bietet. Dafür werden verschiedene Informationsquellen (v.a. einzelne „outputs" wie Publikationen oder Patente, aber auch Angaben zum „esteem" wie z.B. Preise oder die Wahl in bedeutsame Gremien und zum „environment" einer Institution, z.B. Stellen, Drittmittel, Dissertationen) an ein fachlich organisiertes Gutachterpanel gesendet. Aus der Einzelbenotung der heterogenen Informationen bildet dieses Team schließlich eine Gesamtnote, die den Leistungsvergleich zwischen wissenschaftlichen Einheiten (*Units of Assessment*) innerhalb eines Wissenschaftsgebiets ermöglicht.

Hingegen sehen die deutschen und niederländischen Verfahren Einzelfallbegutachtungen von wissenschaftlichen Einrichtungen vor. Entscheidungen werden nicht nur auf Grundlage von Akten, sondern in einem interaktiven Kontext erarbeitet (*Begehung* bzw. *site visit*). Am Ende spricht ein meist interdisziplinär zusammengesetztes Gutachterteam inhaltliche Empfehlungen zur Weiterentwicklung von Instituten aus. In Form einer Checkliste werden die Gutachter/innen dazu angehalten, heterogene Informationen über die Leistung und Leistungsfähigkeit einer je spezifischen wissenschaftlichen Einrichtung in ein Gesamtbild zu integrieren. Es werden jedoch keine Noten vergeben. Die niederländischen und deutschen Verfahren zielen deshalb nicht auf einen institutionellen Vergleich vergangener Leistungen ab, sondern sie tragen deutliche Züge einer Beratung hinsichtlich der zukünftigen Entwicklung einer gesamten wissenschaftlichen Organisation.

Für unsere Frage ist wichtig, dass die jeweiligen Verfahrensregeln zwar bestimmen, welche Informationen erbracht, beurteilt und am Ende kommuniziert werden sollen (numerisch/inhaltlich). Diese Regeln konstituieren die Typen einer vergleichend bewertenden und einer fallspezifisch beratenden Evaluation. Bei der Urteilsfindung selbst verfügen die Gutachter/innen jedoch über eine große Freiheit hinsichtlich der Auswahl, Gewichtung, Ausdeutung und Verwendungsweise von Gütekriterien. Die scientific community ist in den Evaluationsverfahren an zentraler Entscheidungsstelle positioniert und in ihrem Handeln nicht völlig durch Verfahrensregeln bestimmt.

Aus diesen Verfahren haben wir unterschiedliche Institute ausgewählt und deren Evaluationsprozess aus verschiedenen Perspektiven beleuchtet. Neben Dokumenten haben wir Interviews mit den jeweiligen Evaluationsagenturen, Institutsleitungen und -administrationen (vor und nach der Evaluation) und vor allem mit den beteiligten Gutachtern geführt und analysiert. Mit der prozessnahen und multiperspektivischen Erhebung sind wir dem methodischen Problem begegnet, dass uns das Kernereignis, die Kommunikation der Begutachtenden im Panel, verschlossen blieb.[6]

5 Es handelt sich um das Forschungsprojekt der Forschungsgruppe Wissenschaftspolitik am Wissenschaftszentrum Berlin für Sozialforschung: „Urteilsbildung im Peer Review. Internationale Fallstudien zur Evaluation von wissenschaftlichen Einrichtungen". Vgl. auch den Beitrag meiner Kolleginnen Silke Gülker und Dagmar Simon in diesem Band.

6 Mit dem Problem, die Gespräche in Gutachterpanels allenfalls beobachten, aber nicht aufzeichnen zu dürfen und damit diese Kommunikationsdynamik auch nicht rekonstruieren zu können, haben alle mir bekannten Studien zu

Im Zentrum der Interviews standen keine expliziten Deutungen, sondern hinreichend detaillierte Erzählungen über den Verlauf der konkreten Begutachtung, Erläuterungen der Vorgehensweise von Gutachterinnen und Gutachtern anhand konkreter Beispiele sowie Berichte über die dabei aufgetretenen Diskussionen und Probleme. Die Rekonstruktion zielte auf die Grundorientierungen, die Handlungen von Gutachtern zugrunde lagen.

3 Institutionelle Evaluation – Eine neue Handlungssituation?

Kommen wir zur ersten empirischen Teilfrage: Inwiefern entsteht im Kontext solcher institutionellen Evaluationen überhaupt eine neue Handlungssituation für Gutachter/innen? Das ist schon deshalb keine triviale Frage, weil Evaluationsverfahren die konkreten Einzelentscheidungen von Gutachter/innen nicht determinieren, Bewertungen von Forschungsleistungen zum Alltag von Wissenschaftler/innen gehören und bis in die Begriffsverwendung hinein Evaluationen mit dem wissenschaftseigenen Peer Review verschwimmen (z.B. Hirschauer 2002, Neidhard 2010). Welche Veränderungen lassen sich also ausmachen, wenn man die Verfahren institutioneller Evaluationen sowie die Situationsdeutungen von Evaluierten und Gutachtenden betrachtet? Das Verfahren lässt insbesondere drei Verschiebungen sichtbar werden.

In institutionellen Evaluationen sind nicht mehr einzelne Personen, Publikationen oder Forschungsvorhaben der *Begutachtungsgegenstand*, sondern die Gesamtleistung einer ganzen wissenschaftlichen Organisation tritt in den Vordergrund. Damit wandern auch zusätzliche Kriterien und Informationsquellen in die Evaluation ein, die dazu beitragen sollen, ein breites Gesamtbild der jeweiligen Einrichtung zu zeichnen und zu erkennen. Dazu gehören je nach Schwerpunkt des Verfahrens einzelne Publikationen, Listen über den Gesamtoutput, Personal-, Finanzierungs- und Drittmittelbilanzen, aber auch programmatische Selbstbeschreibungen der jeweiligen wissenschaftlichen Einrichtung. Auch wenn es keine verfahrensmäßigen Bestimmungen der Handhabe dieser heterogenen, teilweise neuartigen Informationsquellen gibt, verbinden sich hiermit verschiedene Verdachtsmomente. Das schwerwiegendste ist sicherlich, dass die Orientierung an quantitativen Indikatoren zunehme (z.B. Kieser 2010), um die Fülle an Einzelinformationen überhaupt bewältigen und zu einem Gesamtbild integrieren zu können. Da neben der wissenschaftlichen Qualität auch Fragen der geeigneten Infrastruktur sowie Leitungs- und Organisationsstruktur eine wichtige Rolle spielen, müssen Gutachter/innen zudem über Gegenstandsbereiche jenseits ihrer Fachexpertise urteilen. Es stellt sich deshalb die empirische Frage, welche Informationen für die Gutachter/innen in den jeweiligen Verfahren zentral sind, d.h. mit welchen Selektivitäten sie diesen neuen Gegenstand für sich handhabbar und beurteilbar machen (siehe 5.1).

Eine zweite Verschiebung betrifft den *Begutachtungsprozess*. Während bei der Begutachtung von Publikationen oder Forschungsvorhaben die Initiative stets bei den Begutachteten liegt, man etwas zu gewinnen hat und sich den jeweiligen Wettbewerb (Förderinstitutionen, Publikationsorte und -formate) aussuchen kann, ist das bei institutionellen Evaluationen nicht mehr der Fall. Kein Institut lässt sich freiwillig evaluieren und hat nur wenig Einfluss auf das Evaluationsverfahren, denn es wird turnusmäßig von Wissenschaftsorganisationen zur Teilnahme aufgefordert. Alle Evaluationen stellen damit ohne akuten Anlass den inhaltlichen und finanziellen Fortbestand von Instituten in Frage und untermauern diesen Anspruch mit der Kopplung an Förderentscheidungen. Unter diesen Vorzeichen sind Evaluationen aus Sicht der Institute künstlich erzeugte Krisensituationen, die es wie eine Prüfung möglichst unbeschadet zu überwinden gilt. Empirisch drückt sich das zum einen in der selbstverständlichen Orientierung an einer positiven und problemlosen Evaluierung aus. Zum anderen sind die umfangreichen Vorbereitungshandlungen der Institute zu nennen, die sämtlich darauf abzielen, bereits im Vorfeld alles zu tun, um den Ruf, die Finanzierung und die inhaltliche Ausrichtung der jeweiligen wissenschaftlichen Einrichtung nicht zu gefährden. Das

kämpfen. Vgl. den Beitrag von Thamar Klein und Meike Olbrecht in diesem Band sowie Lamont 2009, Langfeldt 2001, Travis/Collins 1991.

ideale Ziel ist deshalb, von den Gutachter/innen möglichst nicht kritisiert, sondern unterstützt zu werden. Folglich muss man die Vorbereitungshandlungen der Institute als Überzeugungs- oder Kritikvermeidungspraktiken interpretieren, die zu einer Bestätigung führen sollen und je nach Verfahrenstypus variieren. Sie reichen von der Schaffung von Unterstützungsstrukturen zur Produktion von Forschungsoutputs, der Selektion geeigneter Kandidaten für die Präsentationen vor Ort oder der Einreichung von Outputs über das Einkaufen von Stars bis zum aufwendig betriebenen Eventmanagement mit Probeevaluationen. Bei aller Verschiedenheit dieser Praktiken haben sie doch alle das gleiche Ziel: Alles was die Gutachtenden beobachten könnten, wird vorab einer intensiven internen Überprüfung unterzogen. Die wissenschaftsspezifische Präferenz für eine kritische Auseinandersetzung ist in eine öffentlich einsehbare Handlungssituation eingebettet, die für evaluierte Kollegen potentiell existenzgefährdend ist. Ob daraus eine Tendenz zur wenig kritischen Begutachtung im Sinne eines „Nichtangriffpakts" (Schimank 2005: 149) beziehungsweise ein bloßes „akademisches Ritual" (Michaels 2010) folgt, Gutachter/innen ihre Position als „epistemic elites" nutzen und als „arbiters of excellence" partikulare Maßstäbe durchsetzen (Martin/Whitley 2010: 67) oder vielleicht doch gemeinsame Grundorientierungen bei der Urteilsfindung zu beobachten sind, diskutieren wir im Abschnitt 5.2.

Eine dritte Verschiebung betrifft schließlich den *sozialen Kontext der Urteilsfindung* selbst. Statt eines individuellen und anonymen Gutachtervotums steht jetzt ein öffentlich zugängliches und kollektiv getragenes Urteil eines zumeist heterogen zusammengesetzten Gutachterteams im Zentrum. Deshalb stellt sich die Frage, ob das Gutachterurteil durch spezifische Gruppendynamiken präformiert wird und welcher Art diese sind. Stößt man auf Kompromissbildungen zwischen unterschiedlichen Positionen? Reihen sich nur Einzelmeinungen aneinander? Erzeugen erst Verfahrensregeln oder Outputquoten eine Einigung? Oder finden in solchen Gutachtergruppen Normbildungsprozesse auf Grundlage von geteilten Standards statt? (siehe 5.3).

Institutionelle Evaluationen bilden also in mehrerer Hinsicht eine neuartige und spannungsreiche Handlungssituation für die Gutachter/innen. Wie sie auf die Aufforderung an diesen teilzunehmen reagieren, welche Grundorientierungen dabei sichtbar werden und auf welche Weise sie diese neue Herausforderung bewältigen, erörtere ich im Folgenden.

4 Reaktionsweisen und Grundorientierung von Gutachter/innen: Vom wissenschaftspolitischen Auftrag zur professionellen Verpflichtung

Wissenschaftlerinnen und Wissenschaftler reagieren nicht passiv auf neuartige Regulierungsformen, sondern bringen ihre eigenen Orientierungen aktiv ein. Die Frage, warum und wozu Gutachter/innen an in mehrerer Hinsicht spannungsreichen und sehr arbeitsintensiven Evaluationen teilnehmen, ist ein instruktives Beispiel hierfür. Gutachter/innen werden für gewöhnlich von den jeweiligen Evaluationsagenturen angefragt und überlegen nicht lange, ob sie teilnehmen sollen. Sofern die Zeit es zulässt, gilt ihnen die Teilnahme als eine Selbstverständlichkeit.[7] An Evaluationen zu partizipieren wird weder als ein grundsätzlich problematisches noch ablehnbares Unterfangen thematisiert. Die Teilnahme gilt unter den Gutachterinnen und Gutachtern als eine Pflicht gegenüber der scientific community, der man sich nicht entziehen darf: „I think part of my responsibility is not only to do my job in isolation, but to do my job in the context of the community that I relate to, and it's in the interest of that community". Sich in den Dienst der jeweiligen Gemeinschaft zu stellen heißt natürlich nicht, keinen eigenen Vorteil hieraus zu ziehen. So fühlen sich die Gutachter/innen geehrt, diese Rolle zu übernehmen, sie interessieren sich für das Innenleben der evaluierten Einrichtung, können einen Überblick über die Entwicklung und den aktuellen Stand von Forschungsgebieten bekommen,

7 Diese Selbstverständlichkeit tritt in den Interviews dergestalt auf, dass die Gutachter/innen auf die Frage „Wie kam es dazu, dass Sie Gutachter/in wurden?" allenfalls spekulierten, warum Evaluationsagenturen sie angefragt hatten. Ihre Eigenmotivationen blieben dabei aber ausgeblendet und mussten explizit erfragt werden.

sie sind auf die Sicht- und Begründungsweisen der Gutachterkollegen gespannt, können über den Ablauf solcher Evaluationen etwas für die zukünftige Beurteilung der eigenen Institution lernen und schließlich muss die professionelle Selbstkontrolle gewahrt bleiben: „I think it's, since somebody has to do it, it's better that […] you don't refuse to participate". Die Verfolgung genuin wissenschaftlicher Eigeninteressen, die Gewissheit auch über ganze wissenschaftliche Einrichtungen ein Urteil fällen zu können[8] und die Verpflichtung zur ‚community work' setzen Evaluationen in ein Licht, als wären sie eine gewohnte wissenschaftliche Praxis. Ein Ausdruck davon ist, dass die Gutachter/innen die konkreten Evaluationskriterien der Verfahren oftmals nicht präsent haben. Wissenschaftspolitisch initiierte Evaluationen werden also durch die Brille von Wissenschaftler/innen beobachtet, entlang der dort gültigen Regeln, Normen und Wertorientierungen interpretiert und letztlich durch das Einrücken in ihre Handlungsroutinen normalisiert.

Das zeigt sich auch an der spezifischen Interpretation von Gutachter/innen, wozu und für wen Evaluationen dienlich sind. Besonders auffällig ist nämlich die Grundhaltung, Evaluationen weniger als eine Außenkontrolle im Dienst der Wissenschaftspolitik denn als *kollegiale Unterstützung* aufzufassen. Selbst im hierzu weniger geeigneten RAE, das ja einen numerischen Verteilungsschlüssel und keine inhaltlichen Rückmeldungen erarbeitet, findet sich eine solche Primärorientierung: „The first reason (for the RAE) which is probably the only valid one, really is *to help* universities benchmark their research against their competitors". In den beiden anderen Verfahren, tritt sogar ein Selbstverständnis der Gutachtenden als *kollegial beratende Instanz* deutlich hervor: „We are not sitting there to make difficulties, to look for all that's wrong". „It was not as much to find faults but, are they going in the right direction? And are there points of improvement? Can things be done in a better way?"

Die Evaluationsverfahren mit ihren interpretationsbedürftigen Kriterien („Qualität", „Produktivität", „Effektivität", etc.) und die Positionierung von Fachgutachtenden an zentraler Entscheidungsstelle bieten zwar den Raum, professionseigenen Normen und Handlungsweisen zu folgen, sie ziehen aber auch Grenzen. Das kann man gut an den Begutachtungspraktiken beider Evaluationstypen beobachten.

5 Begutachtungspraktiken

Ich konzentriere mich im Folgenden auf ausgewählte Aspekte des Gutachterhandelns, die unmittelbar mit den zuvor angeführten Verschiebungen im Rahmen institutioneller Evaluationen verknüpft sind. Gutachter/innen sind keine Organisationsanalysten und müssen dennoch über die Leistung(sfähigkeit) von wissenschaftlichen Einrichtungen urteilen. Mit welchen Selektivitäten machen sie sich diesen neuen Gegenstand handhabbar? (5.1) Gutachter/innen müssen unter großem Handlungsdruck potentiell folgenreiche Entscheidungen über ihre Kollegen fällen. Welche Standards kommen dabei zum Einsatz und inwiefern genügen sie wissenschaftlichen Gütekriterien? (5.2) Gutachter/innen mit je eigenen Sichtweisen treffen in Panels kollektive Entscheidungen. Unterliegen solche Gruppenentscheidungen speziellen Selektivitäten oder werden nur Einzelmeinungen von Gutachtern hintereinander gestellt? (5.3)

5.1 Handhabung des neuen Gegenstands ‚wissenschaftliche Organisation'

Alle Verfahren fordern von den evaluierten Einrichtungen heterogene Informationen an. Den Gutachtergruppen bleibt aber überlassen, welchen Stellenwert verschiedene Informationsquellen bei der Urteilsfindung haben sollen. Die Verfahren liefern eine Vielzahl möglicher Betrachtungsweisen, aus denen die Gutachter/innen die fach- und fallspezifisch angemessenen wählen. Verfolgt man, welche Informationen die Gutachter/innen aus welchen Gründen selektieren, dann sind die Zuverlässigkeit

8 Eva Barlösius (2008) spricht von einer „Urteilsgewissheit" von Gutachterinnen und Gutachtern selbst bei Fragen, die über ihre Fachexpertise weit hinausreichen.

sowie die Kompatibilität mit eingeübten Begutachtungsweisen und den primär verfolgten Evaluationszielen von zentraler Bedeutung.

Das britische Research Assessment Exercise gewichtet bereits qua Verfahren die drei zentralen Informationsquellen. Das Gesamturteil soll mindestens zu 50% auf „outputs" und jeweils zu 5% auf „environment"- bzw. „esteem"-Indikatoren beruhen. In der Begutachtungspraxis lässt sich eine weitere Konzentration auf den Stellenwert von „outputs" beobachten (zwischen 60 % bei den Ingenieuren und 80% bei den Historikern). Der Gesamtfall ‚Organisation' wird bereits durch dieses Verfahren in Fachbereiche und Einzelinformationen zerlegt und dann in publizierte Einzelakte weiter klein gearbeitet. Erst am Ende werden die Einzelurteile (unqualified, Noten von 1 bis 4) addiert und mit den ebenfalls einzeln bewerteten „esteem"- und „environment"-Informationen zu einem numerischen „quality profile" verrechnet. „Esteem" und „environment" treten in den Hintergrund, weil diese Informationen unter den Gutachter/innen als schwer kontrollier- und einschätzbare Selbstdarstellungen gelten. Sie dienen ihnen jedoch als ein flexibel einsetzbares Korrektiv, z.B. um dem Problem zu begegnen, dass Publikationen nicht in allen Disziplinen den gleichen Stellenwert haben. Im RAE wird der komplexe Gegenstand „wissenschaftliche Einrichtung" durch die Zerlegung in einzelne Fachgebiete, in Einzelleistungen und Einzelurteile handhabbar gemacht. Auf diese Weise bleibt der Eigenwert organisatorischer Fragen begrenzt und die Gutachtenden können auf ihre Erfahrungen aus dem klassischen Peer Review zurückgreifen.

Eine ganz andere Selektivität findet man hingegen bei den niederländischen und deutschen Verfahren. Dort sollen fallspezifische Urteile und Empfehlungen bezüglich einer Gesamtorganisation gefällt werden. Hierfür muss man jedoch den jeweiligen Fall aus verschiedenen Einzelinformationen zunächst konstruieren und zu einem Gesamtbild zusammenfügen und nicht wie im RAE in Einzelakte zerlegen: „You have an impression and you try to articulate that impression and then those criteria are helpful". Deshalb richtet sich die Aufmerksamkeit der Gutachter/innen vor der Begehung insbesondere auf die Programmatiken, Selbstevaluationsberichte und Stärken-Schwächen-Analysen der Gesamtorganisation und Forschungseinheiten. Vor allem dort wird nämlich das Selbstbild der Institution konstruiert, das unter Zuhilfenahme weiterer Informationsquellen von den Gutachter/innen auf seine Konsistenz und Tragfähigkeit überprüft wird. So gibt der Vergleich mit dem vorherigen Evaluationsbericht Aufschluss über die Problemwahrnehmung und Problemlösungskapazität der Institution. Auch die umfangreichen Publikationslisten, (Drittmittel-)Bilanzen oder Zitationsanalysen dienen den Gutachtern nicht so sehr für die direkte Bewertung des Gesamtfalls.[9] Vielmehr sind es flankierende Informationen, die relativ unabhängig von den Selbstbeschreibungen sind und insofern ein eigenständiges Gesamtbild zeichnen, das dann zur Konsistenzprüfung eingesetzt werden kann. Die Glaubwürdigkeit und Angemessenheit der aus Gutachtersicht zentralen „Selbstbeschreibungen mit dem üblichen Selbstlob" werden also ebenso wenig wie im RAE einfach vorausgesetzt. Aber diese Verfahren bieten die Möglichkeit, den Realitätsgehalt und die Angemessenheit solcher Selbstdarstellungen spätestens während der interaktiven Begehungen zu überprüfen und so als analytisches Mittel zu nutzen. Die Gutachter/innen nehmen in diesen Verfahren also tatsächlich die gesamte Einrichtung in den Blick und müssen hierfür ein Gesamtbild generieren. Es steht deshalb der Zusammenhang von einzelnen Informationen im Zentrum und nicht wie im RAE die Zerlegung in einzelne Aspekte. Mit der Konsistenzprüfung von Programmatiken stellen die Gutachter/innen auch hier Informationen ins Zentrum, die sie durchaus gewohnt sind zu bewerten und zu kommentieren. Sie behandeln den Gegenstand Organisation entlang den Inhalten wissenschaftlich relevanter Produkte.

9 Aus der Perspektive der evaluierten Institute mag schon deshalb der Eindruck einer herausragenden Bedeutung der quantitativen Leistungsbemessung entstehen, weil ein großer Teil der Vorbereitung in der Erstellung von Tabellen und Etablierung von Monitoring-Systemen besteht. Dass die Gutachter/innen maßgeblich auf Grundlage dieser Daten urteilen, kann auf Grundlage unserer Erhebung jedenfalls nicht bestätigt werden. Diese quantitativen Informationen helfen schließlich nur wenig in einem Verfahren, das in hohem Maße auf Empfehlungen hinsichtlich der zukünftigen Organisationsentwicklung ausgelegt ist.

5.2 Begutachtungsweisen im Kontext der Verfahren

Wie gehen die Gutachter/innen aber mit den jeweils ausgewählten Primärquellen um und welche Grundorientierungen bringen sich darin zum Ausdruck? Aus der Pflicht gegenüber der scientific community folgt zunächst, dass Gutachter/innen ihre Aufgabe sehr ernst nehmen und einer zweipoligen Haltung folgen: „to make sure we did *the job properly* but that we were also *very fair to the community*". Damit sind einerseits die Einhaltung akzeptabler Standards („proper job") und andererseits die angemessene Anwendung auf die jeweilige Wissenschaftsart („fair job") gemeint.[10] Diese Haltung stellt ein Regulativ dar, damit der Begutachtungsprozess weder in eine unkritische Interessenpoltik für das eigene Fachgebiet noch in überkritische Leistungsanforderungen abgleitet.[11] Die Durchsetzung dieser Grundhaltung trifft allerdings auf verfahrensspezifische Problemlagen.

Mit der Fokussierung auf einzelne „outputs" im Research Assessment Exercise entsteht für die Gutachter/innen zunächst das Problem, viele hundert Einzelbegutachtungen vornehmen zu müssen: „The major challenge for quality was getting the job done!" Deshalb ist auch erwartbar, dass unter Handlungsdruck Beschleunigungs- und Rationalisierungsstrategien entwickelt werden. Aufschlussreich ist dabei, wie abgekürzt wird und aus welchen Gründen heraus. Es wäre naheliegend und im Rahmen des Verfahrens auch möglich, dass die Gutachter/innen ihre Arbeitslast durch die Auswahl nur weniger Outputs, den Verzicht auf eine (zumindest) von zwei Gutachter/innen vorgenommene Bewertung oder durch die Hinzunahme von metrischen Daten bereits publizierter und begutachteter Outputs bewältigten. Genau das geschieht aber nicht, weil die Gutachter/innen auf diese Weise unzulässig abkürzen würden und kein gesichertes Urteil fällen könnten.[12] Die Beschleunigung und Rationalisierung der Einzelbegutachtungen erfolgt vielmehr durch eine Radikalisierung von wissenschaftsspezifischen Deutungsschemata. Das Spezielle solcher Deutungsschemata ist, dass es sich um hochgradig implizite und generalisierte Gesichtspunkte handelt, mit deren Hilfe Outputs durchmustert werden. Ein beschleunigtes Lesen und Interpretieren ist hierfür zwingend erforderlich. Den impliziten Charakter dieser Deutungsschemata kann man sich an dem Phänomen vergegenwärtigen, dass die Gutachter/innen nicht von einhelligen Begutachtungsweisen im Team ausgingen und entsprechend überrascht waren, als sie in sogenannten „calibration sessions" und den nachfolgenden Einzelbewertungen auf eine hohe Übereinstimmung der Urteilsweisen und Urteile stießen. In diesen „calibration sessions" haben die Gutachter/innen ihre impliziten Urteilsweisen explizit gemacht. Dafür hat jeder Publikationen ausgewählt, die aus individueller Sicht das Notenspektrum zwischen ‚unclassified' und der Bestnote 4 repräsentieren. Dann haben alle Panelmitglieder unabhängig voneinander dieses Sample bewertet. Neben der Erkenntnis erstaunlich ähnlicher Urteilsweisen wurde bei der Diskussion von abweichenden Fällen deutlich warum und wer zu harsch oder zu milde urteilt. Die Urteilsweisen von Gutachterinnen und Gutachtern wurden so ausfindig gemacht und eingeordnet: „Very quickly you create an ethos

10 Michèle Lamont et al. (2009a) sprechen von „Fairness as Appropriateness". Damit ist vor allem eine „epistemic contextualisation" gemeint. Diese Norm sorgt dafür, dass epistemologische Vorlieben nicht einfach auf jeden beliebigen Fall projiziert werden, sondern der zu begutachtende Gegenstand über die Angemessenheit von Standards entscheidet.

11 Die Geltung der Norm zeigt sich besonders gut bei Abweichungen. Zum Beispiel wurde von einem schwierigen Gutachterkollegen berichtet, der eine neue Forschungsrichtung vertrat, generell bevorzugte und in seinen Urteilen entsprechend von der Gutachtergruppe abwich. Der Vorsitzende suchte das Gespräch und dieser Gutachter revidierte seine Urteile. Das klassische Beispiel für überzogene Leistungserwartungen ist die Frage danach, was ein wirklich sehr guter und relevanter Beitrag ist. „Paradigm shifting work" und selbst „interesting ideas which change your look on a certain problem" sind nicht erwartbar und gerade in einer Situation kein geeigneter Maßstab, in der eine überkritische Bewertung zu existentiellen Problem führen kann.

12 Interessanterweise gibt es unter den Gutachterinnen und Gutachtern immer wieder den Verdacht, dass zwar nicht im eigenen, aber sicherlich in anderen Panels so vorgegangen würde. Besonders deutlich ist die Ablehnung illegitimer Abkürzungen hinsichtlich einer rein bibliometrischen Bewertung von Outputs, die alle Gutachter/innen aus den bekannten Gründen ablehnen: Weder die Häufigkeit der Zitationen, die Menge oder der Publikationsort gebe Aufschluss darüber, ob es sich im konkreten Fall um einen guten Forschungsbeitrag handele.

for the panel". An den Standards ist interessant, dass sie weder absoluten Maßstäben folgen noch vom jeweiligen Begutachtungsfall ablösbar sind. So könne die Bestnote eigentlich nur an Outputs mit dem Potential zum Paradigmenwechsel – also fast nie! – vergeben werden, so dass es sich stets um ein relatives Urteil über realistisch erwartbare Leistungen handelt. Ebenso ist es unmöglich ein sachhaltiges Urteil zu fällen, wenn man nicht die konkrete Publikation in Augenschein nimmt. Ein beliebtes Beispiel hierfür sind Reviews. Sie können eine einfache Zusammenfassung aktueller Forschungen und damit zwar nützlich, aber ohne wissenschaftlichen Eigenwert sein, selbst wenn sie im Topmagazin Nature veröffentlicht und viel zitiert werden. Umgekehrt ist es aber auch möglich, dass in dem Reviewartikel eine neuartige Frage oder Erklärung generiert wurde und dieser damit äußerst wertvoll ist. Das lässt sich aber nur durch Lesen und Deuten herausfinden.

Ist die grundsätzliche Frage geklärt, ob es sich überhaupt um einen eigenständigen Forschungsbeitrag handelt, dann wechselt die Aufmerksamkeit der Gutachtenden von der Textgattung auf die Textgestalt. Entlang wichtiger Deutungsdimensionen explorieren sie an geeigneten Textstellen die Werthaltigkeit des Beitrags: Was ist der vom Text selbst gesetzte Anspruch? (Überschrift) Was ist daran neu? (Abstract) Stützt die empirische Evidenz den erhobenen Anspruch? (Datengrundlage) Ist die Argumentation klar und konsistent? Welchen verallgemeinerbaren Wert haben die empirischen Einzelergebnisse? (Diskussion). Gelingt dem Text keine gelungene Kommunikation entlang dieser Dimensionen, dann schließen Gutachter/innen auf Schwächen der unternommenen Forschung. Der Begutachtungsprozess bezieht sich also auf den zur Verfügung stehenden Gegenstand, d.h. vor allem auf die Art und Weise wie Forschung kommuniziert wird: „It's all about communication". Durch die Verwendung textbezogener Relevanz, Konsistenz und Stimmigkeitskriterien kann die Begutachtung beschleunigt werden. Eine ganz andere Beschleunigungsform folgt schließlich aus dem Verfahren selbst. Bei den wenigen stark abweichenden Voten ist die Bereitschaft zur Angleichung der vergebenen Noten schon deshalb groß, weil die Bewertung eines einzelnen Outputs für die kumulative Gesamtnote kaum ins Gewicht fällt. Das Gutachterhandeln ist also weiterhin durch wissenschaftliche Normen bestimmt, deren Einhaltung im Kontext eines aufwendigen Verfahrens allerdings immer schwieriger wird. Diese Normen unterbinden (bislang) vielleicht effizientere, aber als illegitim angesehene Abkürzungsstrategien und erzeugen erst den von allen Gutachtern beklagten Zeitaufwand.[13]

Auch bei den niederländischen und deutschen Verfahren spielen allgemeine Deutungsschemata der Stimmigkeit eine wichtige Rolle. Vor allem jedoch auf einer höher aggregierten Ebene als im RAE: Statt Einzelleistungen stehen Gesamtprogrammatiken im Zentrum. Widersprüchlichkeiten, Inkonsistenzen oder fehlende Explikationen geben den Gutachtenden den Anlass für Nachfragen, Kritiken und den Verdacht, dass die Autoren selbst nicht genau wüßten, wofür sie was tun. Sowohl beim Lesen der Unterlagen wie auch der interaktiven Begehung sind die Gutachter/innen auf der Suche nach möglichen Problemlagen in der Ausrichtung und den Arbeitszusammenhängen des Instituts. Das ist natürlich in einem prüfungsartigen Evaluationskontext besonders schwierig, weil dort ja gerade nicht von einer freimütigen Offenlegung von bearbeitungswürdigen Problemlagen ausgegangen werden kann. Die Darstellung der Erfolge steht für die Institute im Vordergrund und führt gelegentlich zu einem institutsinternen Wettbewerb darüber, wer präsentieren darf.

Die Gutachter/innen sind darauf gefasst, hinter die mit allerlei rhetorischen, ästhetischen und inszenatorischen Mitteln aufgebaute Fassade solcher Selbstbeschreibungen gelangen zu müssen, um urteilen zu können. Die Stärken-Schwächen-Analysen und die Umgangsweise mit den Empfehlungen der letzten Evaluation bieten den Gutachterinnen und Gutachtern einen Zugang, um einen ersten Einblick in die Problemsicht und Problembearbeitungsweisen der jeweiligen Einrichtung zu bekommen. Sie begeben sich aber noch weiter auf Indiziensuche und verwenden im Wesentlichen

13 Im nächsten RAE, Research Excellence Framework (REF) genannt, soll die Verwendung von metrischen Daten wichtiger werden. Ob die Gutachter/innen trotz massiver Kritik aus pragmatischen Gründen dennoch auf sie zurückgreifen, bleibt abzuwarten.

drei Strategien, um die sachliche Angemessenheit, Glaubwürdigkeit und Realisierbarkeit von Institutsprogrammatiken zu überprüfen. Erstens werden die Selbstdarstellungen an zusätzlichen Informationen gespiegelt, die teilweise zur Verfügung gestellt werden (Outputs, Infrastruktur, Finanzierung, Kooperationsbeziehungen), vor allem aber den Gutachter/innen „auch ohne die Unterlagen" vorliegen, weil sie auf ein „gewisses Vorwissen" rekurrieren können. Dieses Vorwissen ist ein Erfahrungswissen, das zum Beispiel die Besonderheiten des wissenschaftlichen Themenbereichs, die institutionelle Einbettung des zu begutachtenden Instituts oder typische Schwierigkeiten des Wissenschaftsbetriebs umschließt. Es handelt sich dabei also weniger um ein spezialisiertes und formalisiertes Ableitungswissen, als viel mehr um eine generelle Kenntnis des Handlungsfeldes mit seinen typischen Herausforderungen und Problemlagen: „Irgendwie weiß man das".[14] Zweitens erfolgt eine textimmanente Konsistenzprüfung: „Die Papiere müssen in sich logisch sein oder ne gewisse, einen Zusammenhang aufweisen". Bereits im Studium der Akten suchen die Gutachtenden neben Widersprüchlichkeiten auch nach Indizien für und wider die Glaubwürdigkeit der Selbstdarstellungen, die dann in der direkten Interaktion eine weitere Überprüfung erfahren. Denn „die Papiere [müssen] mit dem was die Leute sagen in Übereinstimmung stehen, so dass es nicht auseinanderfällt". Die Gutachter/innen kommen also bereits mit einem ersten Bild im Kopf in die Institute, das dort bestätigt, verfeinert, ausgebaut oder relativiert, aber nach eigener Auskunft nur selten verworfen wird.

Eine dritte Strategie besteht schließlich darin, die spontanen Reaktionsweisen (insbesondere der Institutsleitung) auf Fragen genau zu beobachten und Schlussfolgerungen daraus zu ziehen. Wie in einer therapeutischen Situation oder in juristischen Verfahren kommt es dabei weniger darauf an, was den jeweiligen Fall betreffend gefragt wird, sondern wie die Antworten ausfallen. Hierzu ein Gutachter:

> *„Also das ist eben die Art, antworten die Leute auf Fragen, die man ihnen stellt, auch auf kritische Fragen, und wie gehen sie mit diesen Fragen um? Weichen sie denen aus, beantworten sie die gar nicht, beantworten sie die klar? Wenn sie die klar beantworten und auch ein Problem eingestehen, ist das im Prinzip schon mal ein Indiz, dass es in die richtige Richtung geht. Und wenn sie ein realistisches Selbstbild auch haben, wie sie sich selber einschätzen, ist das auch ein positives Indiz. Also das sind Indizien, die was mit Glaubwürdigkeit zu tun haben."*

Die Handlungsweise der Gutachter/innen geht weit über das Muster fachlicher Expertise und Kritik hinaus. Vor allem fällt die Nähe zu dem auf, was man in der Soziologie den Bereich professionellen Handelns nennt und in diesem Beispiel die Form des beratenden oder supervisorischen Handelns annimmt. Die Bewertung steht nicht für sich, sondern soll Verbesserungen anregen. Dafür ist aber konstitutiv, dass Probleme offengelegt und kommuniziert werden. Erst vor diesem Hintergrund wird verständlich, dass Gutachter/innen das Eingestehen von Problemen, ein realistisches und problembewusstes Selbstbild der Institute, honorieren, obwohl in der Handlungssituation Evaluation gerade nicht damit zu rechnen ist. Das kritische kollegiale Gespräch über Problemlagen und mögliche Lösungen ist eine wissenschaftsintern besonders anschlussfähige Deutung von Evaluationen, der aber eine verfahrensbezogene Deutung als Prüfung entgegensteht.[15]

An beiden Beispielen von Gutachterpraktiken im Kontext institutioneller Evaluationen müsste deutlich geworden sein, dass Eigenregulierungen der Wissenschaft nicht einfach ausgehebelt werden und noch immer das Handeln strukturieren. Die Verfahren erschweren dies allerdings in unterschiedlicher Weise. Sie generieren einen Zeitdruck, erzwingen Abkürzungsstrategien oder erzeugen eine Prü-

14 Im juristischen oder ärztlichen Handeln entspricht dem ein typologisches Wissen über Standardfälle, Symptomatiken, Problem- oder Motivkonstellationen, mit denen ein konkreter Fall kontrastiert und in seiner Spezifik erschlossen werden kann.

15 So tituliert ein Institutsdirektor seinen Erfahrungsbericht mit der Evaluation der Leibnizgemeinschaft mit „Die Prüfung als Chance begreifen" und betont den „Dialog-Charakter" gegenüber der „Kontrollvisite". Vgl. Leibniz-Journal 3/4, 2006, S. 30f.

fungssituation, die der Verständigung über Verbesserungsmöglichkeiten entgegen steht.

5.3 Gruppendynamik

Kommen wir zur letzten Verschiebung der Handlungssituation von Gutachterinnen und Gutachtern im Kontext institutioneller Evaluationen. Im Unterschied zum klassischen Peer Review fällen die Gutachter/innen ihr Urteil in einer Gruppe. Genauer gesagt müssen Einzelurteile, die Gutachter/innen bei der Durchsicht von Unterlagen, Outputs oder bei der Begehung gefällt haben, von der gesamten Gruppe getragen und gemeinsam nach Außen vertreten werden. Heterogen besetzte Gutachtergruppen stehen also unter einem erhöhten Konsenszwang. Dieses Problem wird in den Verfahren unterschiedlich bewältigt. Im RAE erfolgt eine Abstimmung der Bewertungsweise in den bereits genannten „callibration sessions" vor den Einzelbegutachtungen, es kommt zu Diskussionen zwischen den (meist) zwei für einen Output verantwortlichen Gutachtern, sofern die Urteile zu weit auseinander gehen, und es gibt Konsistenzkontrollen der Urteilsweisen von Einzelgutachtern während des Prozesses. Diese bestehen darin, dass die Varianz der von Gutachtern vergebenen Noten überprüft und ggf. korrigiert wird (siehe das Beispiel in FN 11). Bei der Bildung des Gesamtergebnisses bedarf es keiner weiteren Koordination, da die Einzelnoten nur zusammengerechnet werden. In den niederländischen und deutschen Verfahren werden die Einzeleindrücke der Gutachter/innen im Verlauf der Begehung immer wieder in Gesprächen der Gutachtergruppe gesammelt. Besonders wichtig ist das erste Treffen am Vorabend der Begehung und das letzte, bevor eine Rückmeldung an die evaluierte Einrichtung gegeben wird. Die erste Ab-stimmung gibt vor, ob es sich nach Lektüre der Unterlagen um eine problematische oder unproblematische Evaluation handelt. In der letzten wird selektiert, welche Kritikpunkte und konkreten Empfehlungen von der Gutachtergruppe letztlich vertreten werden. Oftmals werden die als Checklisten vorliegenden Kriterienkataloge erst hier herangezogen, um die gesammelten Eindrücke zumindest im Bewertungsbericht ans Verfahren anzugleichen.

In beiden Evaluationstypen finden sich also Hinweise darauf, dass Gruppenurteile nicht auf das Hintereinanderschalten von Einzelperspektiven der Gutachter/innen reduziert werden können. Beschreibungen wie „ethos for the panel" und der Rückgang auf sehr allgemeine Deutungsschemata verweisen vielmehr darauf, dass in diesen Gruppen grundsätzliche Normen und Wertorientierungen wissenschaftlichen Handelns aktualisiert, eingefordert oder sogar erst gebildet werden, um über Sonderperspektiven hinaus zu gelangen. Die Orientierung an abstrakten und kaum operationalisierbaren Standards macht das Gutachterhandeln zwar wenig berechenbar, aber dennoch zweckrational: Sie ermöglichen eine *Beschleunigung der Urteilsfindung*, weil eine Reduktion auf wesentliche Aspekte erfolgt und nur strittige Sachverhalte diskutiert werden. Die Notwendigkeit, Einwände gegenüber wissenschaftlichen Kollegen begründen und rechtfertigen zu müssen, hat zudem eine *disziplinierende Wirkung*. Denn jeder Gutachtende wird in diesen Gruppen zugleich selbst begutachtet.[16] Zu hart oder zu milde Urteilende, zurückhaltende oder viel diskutierende, sach- oder selbstbezogene Gutachter/innen werden sichtbar. Daran schließen *diskursive Effekte* an, denn der Wert dieser Beiträge bemisst sich daran, ob sie Anschlussfähigkeit in der Gruppe erlangen. Kritische Einwände oder Sonderperspektiven müssen sich nämlich in der Gutachtergruppe bewähren und werden nicht einfach übernommen. Sie finden ihren Weg in das Gesamturteil nur, wenn sie von anderen Gutachter/innen aufgegriffen und damit bekräftigt werden.[17]

16 Im „Forum: Begutachtung in der soziologischen Drittmittelforschung" auf dem Kongress der Deutschen Gesellschaft für Soziologie 2010 haben die sog. Fachkollegiaten, die in ihrem Gremium ebenfalls Gruppenentscheidungen treffen, explizit erwähnt, dass sie ihre Funktion in der Begutachtung der Gutachter/innen sehen und dieser vornehmlich durch Normbildungsprozesse innerhalb der Gruppe hinsichtlich der Angemessenheit von Gutachten nachkommen.

17 Das kann dann auch dazu führen, dass in Begutachtungen ein vergleichsweise singulärer Aspekt herausgegriffen wird, weil dieser allen Gutachtern aufgefallen ist. Sofern ein Fachexperte unter den Gutachtenden ist, hat dieser erhöhte Durchsetzungschancen, weil dieser den Wert eines Beitrags seinen Gutachterkollegen besser begründen

Im Anschluß an Michèle Lamonts These, dass bei Gruppenevaluationen kaum formalisierbare „Customary Rules“ (2009) und die fallbezogene „Appropriateness“ (2009a) des Urteils eine zentrale Bedeutung haben, habe ich zu zeigen versucht, dass sehr allgemeine und nicht mechanistisch anwendbare Deutungsschemata der Konsistenz, Stimmigkeit und Fallangemessenheit auch unter den spezifischen Rahmenbedingungen institutioneller Evaluationen operieren. Diese scheinen gerade in heterogenen Gremien sogar eher wichtiger zu werden und sind sicherlich nicht durch die formalen Verfahrensregeln erzeugt worden.

6 Fazit

Was kann man aus unserem Ausflug in die Innenwelt von Evaluationsprozessen über die institutionellen Folgen von Evaluationen lernen? Ich habe an einer rein institutionalistischen Perspektive kritisiert, dass formale Rahmenbedingungen nicht einfach das empirisch beobachtbare Handeln von Gutachterinnen und Gutachtern bestimmen. Diese agieren auf Grundlage eigener Regeln der Angemessenheit, die keineswegs vom formalen Verfahren erzeugt wurden. Welche Zwecke Verfahren der institutionellen Evaluation auch immer verfolgen, sie sind in ihrer Durchführung zumindest so lange mit wissenschaftsinternen Relevanzen durchdrungen, wie Angehörige der scientific community an zentraler Entscheidungsposition platziert sind. Vor allem geben solche Verfahren an, was begutachtet und in welcher Form Ergebnisse kommuniziert (numerisch/inhaltlich) werden sollen. Wie Gutachter/innen begutachten sollen bleibt jedoch unbestimmt, so dass sie hinreichend Freiraum haben, die Verfahren an wissenschaftliche Standards anschlussfähig zu halten.

Das habe ich an drei wesentlichen Neuerungen, die institutionelle Evaluationen mit sich bringen, gezeigt. Auch wenn dort wissenschaftliche Organisationen beurteilt werden, so stehen mit der Konsistenz und Innovativität von Forschungsprogrammen oder Forschungserzeugnissen weiterhin wissenschaftsspezifische Aspekte im Zentrum. Organisation wird gewissermaßen auf die Frage begrenzt, ob vorhandene Regelungen geeignet sind, interessante Forschung zu unterstützen. Mit diesem neuen Gegenstand geht allerdings das Problem einher, dass eine Vielzahl heterogener Informationen von den Gutachtern durchgemustert werden müssen und ihr Urteil für die Zukunft einer ganzen wissenschaftlichen Einrichtung folgenreich sein kann. Gutachter/innen stehen damit unter erhöhtem Handlungsdruck, tragen große Verantwortung und haben einen erheblichen Einfluss. Die gerne daraus abgeleiteten Tendenzen zu schematischen, milden oder partikularistischen Urteilsweisen konnten allerdings nicht bestätigt werden. Vielmehr stößt man unter den Gutachtern auf eine Haltung, kollegiale Unterstützung über eine Fall angemessene Kritik auszuüben, die sich an abstrakten Fächer- oder Gegenstandsgrenzen übergreifenden Interpretationsschemata orientiert. Diese sind gerade in heterogen besetzten Gutachtergruppen ein wichtiges Mittel, um zu einem gemeinsamen Urteil zu gelangen. Die paradoxe Folge extern initiierter Überprüfungen im Rahmen institutioneller Evaluationen könnte also sein, dass wissenschaftseigene Kriterien radikalisiert und auf neue Gegenstandsbereiche ausgedehnt werden.

kann.

Literatur

Barlösius, Eva, 2008: Urteilsgewissheit und wissenschaftliches Kapital, in: *Matthies, Hildegard/ Simon, Dagmar (Hg.):* Wissenschaft unter Beobachtung: Effekte und Defekte von Evaluationen. Wiesbaden: VS Verlag, 149-196.

Gläser, Jochen / Lange, Stefan / Laudel, Grit / Schimank, Uwe, 2008: Evaluationsbasierte Forschungsfinanzierung und ihre Folgen, in: *Mayntz, Renate et al. (Hg.):* Wissensproduktion und Wissenstransfer. Wissen im Spannungsfeld von Wissenschaft, Politik und Öffentlichkeit. Bielefeld: Transcript, 145-170.

Gläser, Jochen / Laudel, Grit, 2007: Evaluation without Evaluators: The impact of funding formulae on Australian University Research, in: *Whitley, Richard / Gläser, Jochen (eds.):* The Changing Governance of the Sciences: The Advent of Research Evaluation Systems. Dordrecht: Springer, 127-151.

Hirschauer, Stefan, 2002: Expertise zum Thema „Die Innenwelt des Peer Review. Qualitätszuschreibung und informelle Wissenschaftskommunikation in Fachzeitschriften." Förderinitiative »Wissen für Entscheidungsprozesse – Forschung zum Verhältnis von Wissenschaft, Politik und Gesellschaft« des BMBF. Online: http://www.sciencepolicystudies.de/dok/expertise-hirschauer.pdf.

Kieser, Alfred, 2010: Unternehmen Wissenschaft? Leviathan. Berliner Zeitschrift für Sozialwissenschaft 38, 347-367.

Lamont, Michèle, 2009: How Professors Think: Inside the Curious World of Academic Judgment. Cambridge: Harvard University Press.

Lamont, Michèle / Mallard, Grégoire / Guetzkow, Joshua, 2009a: Fairness as Appropriateness: Negotiating Epistemological Differences in Peer Review. Science, Technology & Human Values 34, 573-606.

Langfeldt, Liv, 2001: The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome. Social Studies of Science 31, 820-841.

Leisyté, Liudvika / Boer, Harry de / Enders, Jürgen, 2010: Mediating Problem Choice: Academic Researchers' Responses to Changes in their Institutional Environment, in: *Whitley, Richard / Gläser, Jochen / Engwall, Lars (eds.):* a.a.O, 266-290.

Martin, Ben / Whitley, Richard, 2010: The UK Research Assessment Exercise. A Case of Regulatory Capture?, in: *Whitley, Richard / Gläser, Jochen / Engwall, Lars (eds.)*, a.a.O, 51-80.

Michaels, Axel, 2010: Rituale der Forschungsevaluation: Die große Begehung der Mittelbaustelle. Frankfurter Allgemeine Zeitung, 15. August 2010.

Neidhardt, Friedhelm, 2010: Selbststeuerung der Wissenschaft: Peer Review, in: *Simon, Dagmar / Knie, Andreas / Hornbostel, Stefan (Hg.):* Handbuch Wissenschaftspolitik. Wiesbaden: VS Verlag, 280-292.

Schimank, Uwe, 1995: Für eine Erneuerung der institutionalistischen Wissenschaftssoziologie. Zeitschrift für Soziologie 24, 42-57.

Schimank, Uwe, 2005: Die akademische Profession und die Universitäten: "New Public Management" und eine drohende Entprofessionalisierung, in: *Thomas Klatetzki /Veronika Tacke (Hg.):* Organisation und Profession. Wiesbaden: VS-Verlag, 143-164.

Torka, Marc, 2009: Die Projektförmigkeit der Forschung. Baden-Baden: Nomos.

Travis, G.D.L. / Collins, H.M., 1991: New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System. Science, Technology, & Human Values 16, 322-341.

Weingart, Peter, 2005: Das Ritual der Evaluierung und die Verführbarkeit der Zahlen, in: *Ders. (Hg.):* Die Wissenschaft der Öffentlichkeit: Essays zum Verhältnis von Wissenschaft, Medien und Öffentlichkeit. Weilerswist: Velbrück, 102-122.

Whitley, Richard / Gläser, Jochen / Engwall, Lars (eds.), 2010: Reconfiguring Knowledge Production. Changing Authority Relationships in the Sciences and their Consequences for Intellectual Innovation. Oxford: Oxford University Press.

Silke Gülker, Dagmar Simon

Nach der Evaluation ist vor der Evaluation? Institutionelle Folgen von Forschungsbewertungen im internationalen Vergleich

1 Einleitung

Was passiert in einem Forschungsinstitut vor und nach einer Evaluation? Welche institutionellen Effekte sind mit diesem Instrument verbunden, das zumeist mehrfache Zielsetzungen gleichzeitig verfolgt: Evaluationen sollen einerseits Informationen liefern für Allokationsentscheidungen – und dies in einem in der Regel harten Wettbewerb um knappe Mittel. Andererseits ist mit Evaluationen stets auch der Anspruch einer Qualitätsentwicklung verbunden.

Im Rahmen eines Forschungsprojektes[1] in der Forschungsgruppe Wissenschaftspolitik am Wissenschaftszentrum Berlin für Sozialforschung (WZB) haben wir uns mit den institutionellen Effekten von Evaluationen im Sinne von forschungsstrategischen Antizipations- und Anpassungseffekten befasst. Über derartige Effekte wird derzeit viel debattiert und zum Teil auch spekuliert. Angenommen werden falsche Anreize durch Evaluationen, die die intrinsische Motivation der Wissenschaftler/innen untergraben, zu Fehlsteuerungen und (nicht intendierten) Anpassungseffekten führen würden (Heintz 2006, Frey 2008; Kieser 2010). Durch verstärkte Anwendungsorientierung gerate etwa die Grundlagenforschung unter Druck, durch einseitige Betonung von Publikationen in refereed journals als Qualitätsmerkmal würde vor allem „nicht riskante" Forschung gefördert. „Numerische Kriterien zur Leistungsbeurteilung eines Forschers begünstigt die Vielschreiberei", wie es jüngst der DFG-Präsident Matthias Kleiner (vgl. Boni 2010) ausdrückte. Die zusätzliche Jagd nach Drittmitteln als zweites wichtiges Evaluationskriterium verändere darüber hinaus Produktionszyklen der wissenschaftlichen Arbeit und ihrer Organisationsformen.

Effekte dieser Art lassen sich nur über langfristige Wirkungsanalysen nachvollziehen und unser Projekt ist nicht dafür geeignet, diese Thesen zu bestätigen oder zu verwerfen. Wohl aber können wir auf der Grundlage umfassender Interviewanalysen zum aktuellen Zeitpunkt nachvollziehen, ob und inwiefern durch Evaluationen organisationale und strukturelle Veränderungen in den Instituten angestoßen werden.

Ausgehend von der Hypothese, dass Verfahrensregeln einer Evaluation einen Unterschied dafür machen, welche Effekte in den Instituten zu erwarten sind, steht ein Vergleich der Verfahrenstypen im Vordergrund. Mit der Evaluation der Leibniz-Gemeinschaft in Deutschland, dem Standard Evaluation Protocol in den Niederlanden und der Research Assessment Exercises in Großbritannien wurden Verfahren ausgewählt, die zum Teil nach grundlegend unterschiedlichen Prinzipien arbeiten.

Im Folgenden werden zunächst die Charakteristika dieser Verfahren kurz erläutert (Abschnitt 2) und daraufhin analysiert, was jeweils in den Verfahrenskontexten als „gute Wissenschaft" definiert ist und wie diese von den Instituten präsentiert wird (Abschnitt 3). Schließlich wird überprüft, welche Indizien darauf hinweisen, ob und wie sich wissenschaftlicher Alltag verändert (Abschnitt 4).

2 Die Verfahren im Überblick

Die drei Verfahren lassen sich zwei Grundtypen zuordnen. Bei Evaluationen der Leibniz-Gemein-

1 Urteilsbildung im Peer Review – Internationale Fallstudien zur Evaluation von wissenschaftlichen Einrichtungen. An diesem Projekt waren Silke Gülker, Sandra Matthäus, Marc Torka und Dagmar Simon beteiligt. Vgl. auch der Beitrag von Marc Torka in diesem Band.

schaft und nach dem Standard Evaluation Protocol steht eine Begehung des zu begutachtenden Forschungsinstituts im Zentrum, die einen interaktiven Prozess zwischen Evaluatoren und Evaluierten erlaubt. Bei der Research Assessment Exercise dagegen wird allein nach Schriftlage bewertet. Damit verbunden sind jeweils unterschiedliche Vorgaben für die Vorbereitung der Institute, für den Prozess der Bewertung durch Gutachter/innen sowie für die Darstellung und Nutzung der Ergebnisse – wie im Folgenden skizziert wird.

Die **Leibniz-Gemeinschaft** wurde 1997 als Nachfolgeorganisation der „Wissenschaftsgemeinschaft Blaue Liste" gegründet und ihr gehören derzeit 86 Forschungsinstitute an, die von Bund und Ländern gemeinsam finanziert werden. Jedes Mitgliedsinstitut der Leibniz-Gemeinschaft wird in der Regel[2] alle sieben Jahre evaluiert. Für das Bewertungsverfahren wird vom Senatsausschuss Evaluierung (SAE) ein Gremium von Fachgutachter/innen zusammengestellt, deren Mitglieder nach den Verfahrensregeln keinen Kooperationskontakt zu dem Institut haben dürfen und gegen die keine anderweitigen Befangenheitsbedenken bestehen.

Kern des Bewertungsprozesses ist ein Vor-Ort-Besuch des Instituts, die so genannte „Begehung". Zu deren Vorbereitung werden den Fachgutachter/innen Unterlagen über die zu evaluierende Einrichtung zur Verfügung gestellt, die nach einem vorgegebenen Frageraster strukturiert sind. Die Unterlagen enthalten Angaben zu den Aufgaben des Instituts, seiner Größe und Struktur, zur Mittelausstattung, zu den Arbeitsschwerpunkten sowie zu den zentralen Arbeitsergebnissen (Publikationen, Konferenzen, Drittmittel). Die Begehung dauert eineinhalb Tage und beinhaltet eine Präsentation der Institutsleitung gegenüber der gesamten Begutachtungsgruppe, arbeitsteilig organisierte Abteilungsbesuche, ein Gespräch mit den Mitarbeiter/inne/n unter Ausschluss der Führungskräfte, ein Gespräch mit der Verwaltungsleitung und Gespräche mit Kooperationspartnern. Neben den Fachgutachter/inne/n sind an der Begehung auch Vertreter/innen von Bund und Ländern beteiligt. Das Ergebnis der Evaluation wird in Form eines ausführlichen Bewertungsberichts, der auch Handlungsempfehlungen beinhaltet, sowohl den Zuwendungsgebern als auch dem Institut übermittelt.

Das **Standard Evaluation Protocol for Public Research (SEP)** wurde im Jahre 2003 gemeinsam von der Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), der Vereiniging van Universiteiten (VNSU) und der Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) verfasst und bildet seither die Grundlage für die alle sechs Jahre stattfindende Evaluation der öffentlich finanzierten Forschung(seinrichtungen) (KNAW et al. 2003). Das Verfahren entspricht in vielerlei Hinsicht dem der Leibniz-Gemeinschaft, weist allerdings manche bemerkenswerte Besonderheiten auf. So wird mit dem SEP etwa nicht allein die wissenschaftliche Leistung, sondern auch die Qualität des Institutsmanagements (Führung, Strategie, Instrumente, Forschungsorganisation) bewertet. Für die Evaluation der Forschungseinrichtungen sind insgesamt vier Kriterien vorgesehen: Qualität, Produktivität, Relevanz sowie Veränderungsfähigkeit einer Organisation und deren Vermögen, realisierbare Strategien zu entwickeln. Das zu evaluierende Institut wird auf einer Skala von 5 (excellent) bis 1 (unsatisfactory) nach den genannten Kriterien bewertet.

Ausgangspunkt der Bewertung der einzelnen Forschungseinrichtungen ist ein Selbstevaluationsbericht, den die Institute selbst zu erstellen haben. Die Selbstevaluation (mid-term) soll jeweils drei Jahre vor oder nach der externen Evaluation stattfinden und einerseits die Umsetzung der Empfehlungen der letzten Evaluation überprüfen sowie andererseits die kommende Evaluation vorbereiten. Außerdem sind die Forschungseinrichtungen angehalten, ein Monitoringsystem zu etablieren, das auch online verfügbar ist und im jährlichen Rhythmus wesentliche Kennziffern über die Entwicklung des Instituts bereitstellt.

Die Begehung (site visit) des Instituts dauert ein bis zwei Tage und beginnt wie in dem Verfahren

2 Das Ergebnis eines Bewertungsverfahrens kann allerdings vorsehen, dass eine nächste Evaluation bereits früher stattfinden muss.

der Leibniz-Gemeinschaft in der Regel mit einer nichtöffentlichen Sitzung der Gutachter/innen. Der weitere Ablauf der Begehung kann allerdings stark variieren. Anders als im Falle der WGL wird kein einheitliches Protokoll umgesetzt, sondern in einem Gespräch zwischen KNAW und dem zu begutachtenden Institut ein paar Monate vor dem Termin der genaue Zeit- und Ablaufplan entwickelt.

Die **Research Assessment Exercises** fanden 2008 zum sechsten Mal statt. Mit jedem Durchgang wurde das Verfahren verändert (vgl. Martin/Whitley 2010). Getragen und organisiert wird die RAE durch den Higher Education Funding Council for England (HEFCE)[3] und die Aufgabe ist, eine Verteilungsgrundlage für den Anteil institutioneller Forschungsförderung zu schaffen. Die RAE sieht vor, dass jede Universität disziplinäre Forschungseinheiten definiert und jeweils eine Bewerbungsunterlage entwickelt. Sie enthält neben einer festgelegten Anzahl an Daten zur Institution und einer Beschreibung der Forschungsstrategie und Organisationsstruktur die Benennung von sogenannten „forschungsaktiven Wissenschaftler/inne/n", für die als Grundlage der Bewertung jeweils vier Publikationen zu benennen sind. Bewertet werden die Unterlagen in disziplinär organisierten Bewertungsgremien. 2008 haben 67 sogenannte sub-panels gearbeitet und zusätzlich waren 15 übergeordnete Gremien (main panels) dafür zuständig, die Konsistenz der Bewertung innerhalb der Fachcluster zu überprüfen. Die Auswahl der Persönlichkeiten zur Besetzung dieser Gremien erfolgt durch ein öffentliches Vorschlagsverfahren: Fachorganisationen und Einrichtungen, die zum Kreise der Nutzer von Wissenschaft zu zählen sind, haben das Recht, Personen vorzuschlagen, die dann in die engere Wahl kommen.

Die Bausteine der Bewerbungsunterlagen sind in den Fachgremien jeweils unterschiedlich gewichtet, den größten Ausschlag für das Resultat gibt aber immer die Bewertung der eingereichten Publikationen. Die Gremien tagen in der Regel fünf Mal im Laufe des Jahres. Zum Ende des Prozesses sind sie aufgefordert, für jede zu bewertende Einheit eine Note zwischen 1* und 4* zu vergeben. Diese Notenskala wird in eine Formel zur Berechnung des Budgets für die nächsten sieben Jahre übertragen und die Resultate aller Forschungseinheiten werden online publiziert.

3 Was gilt als „gute Wissenschaft" und wie wird sie präsentiert?

Die beiden unterschiedlichen Verfahrenstypen – Leibniz-Gemeinschaft und SEP auf der einen Seite und RAE auf der anderen – setzen unterschiedliche Schwerpunkte bei der Definition dessen, was als „gute Wissenschaft" bezeichnet wird. Beim deutschen und niederländischen Verfahren steht die Forschungseinrichtung im Zentrum der Bewertung, in Großbritannien die individuelle Leistung jeweils ausgewählter Wissenschaftler/innen.

Die Leitlinien der Leibniz-Gemeinschaft stellen etwa die Qualität der Forschung des Instituts in den Mittelpunkt, das Forschungsprogramm soll an Kriterien wie „State of the Art" und „Kohärenz" gemessen werden. Darüber hinaus soll die überregionale Bedeutung des Instituts eingeschätzt werden sowie auch die Attraktivität der Einrichtung für Partner und Nutzer. Zusätzlich geht es um eine Bewertung der Institutsstruktur, etwa die Effizienz, Leistungsfähigkeit der Infrastruktur und die Personalentwicklung. Im Falle des SEP wird dies konkretisiert, hier soll auch die Qualität der Institutsleitung bewertet werden. Für die Bewertung der Forschungseinrichtungen sind – wie schon ausgeführt – insgesamt im SEP vier Kriterien vorgesehen: Qualität, Produktivität, Relevanz sowie Dynamik und Realisierbarkeit. Sowohl die Institutsleitungen als auch die Gutachtergruppe haben es also mit einer überaus komplexen Aufgabe zu tun – Inhalt und Prozess einer Forschungseinrichtung müssen von den Instituten so aufbereitet werden, dass sie von den Gutachter/inne/n

3 HEFCE wurde 1992 als Nachfolgeorganisation des University Grants Council gegründet und ist für die Verteilung institutioneller Forschungsförderung und damit für eine der zwei Säulen im Fördersystem zuständig. Der andere Teil wird über die Research Councils projektbezogen vergeben (HEFCE 2007a). In Bezug auf die RAE übernimmt der HEFCE auch Managementaufgaben für die Förderorganisationen von Schottland, Wales und Nordirland.

innerhalb einer knappen Vorbereitungs- und Begehungszeit bewertbar werden.

Gute Wissenschaft in der RAE meint in allererster Linie gute Publikationen. Die Leitlinien des Verfahrens konkretisieren die Gütekriterien eines Papiers wesentlich im Zusammenhang mit dessen Wirkung in der Fachgemeinschaft. Die Bestnote 4* soll an ein Papier vergeben werden, wenn „It has become or is likely to become, a *primary point of reference* in its field or sub-field". Für die zweitbeste Note 3* gilt: „It has become or is likely to become a *major point of reference* in its field or sub-field", für 2*: „It has made or will make *a contribution* to its field or sub-field" und für 1*: „It has made or will make *a limited contribution* in its field or sub-field." Darüber hinaus orientiert sich die Beschreibung guter Wissenschaft an den Begriffen „originality", „significance" und „rigour" und damit an den klassischen Kriterien im Peer Review von Zeitschriftenaufsätzen (HEFCE 2005).

Zu dieser Primärorientierung – Bewertung der Leistungen der Organisation bei Leibniz-Gemeinschaft und SEP auf der einen und Publikationsbewertung der Wissenschaftler/innen bei der RAE auf der anderen Seite – lassen sich allerdings Brüche aufzeigen. So spielen beispielsweise in den Verfahren der Leibniz-Gemeinschaft standardisierbare Leistungen ebenfalls und zunehmend eine wesentliche Rolle in den Vorbereitungsunterlagen. Im Rahmen des umfangreichen Fragekatalogs werden die Publikationsoutputs detailliert und nach Publikationsort klassifiziert abgefragt. Umgekehrt sehen die Leitlinien der RAE seit 2008 eine Bewertung der übergreifenden Forschungsstrategie vor. Zwar war dazu auch in früheren Verfahren ein Textdokument gefordert, erstmals muss dieses Dokument nun aber ebenfalls mit einer Bewertung versehen werden und die Gremien entscheiden vorab, mit welchem Anteil diese Bewertung in die Gesamtnote der Forschungseinheit eingehen soll. In beiden Verfahrenstypen ist also, wenn auch mit sehr divergierenden Schwerpunkten, beides gefragt: Individueller Forschungsoutput und Konsistenz der Einrichtung.

Die unterschiedlichen Schwerpunktsetzungen in den Kriterien werden in den Vorbereitungen der Institute auf die Evaluierung antizipiert. Während der Fokus bei Instituten der Leibniz-Gemeinschaft und der KNAW darauf liegt, das Institut als Ganzes zu präsentieren und hier insbesondere auch in der Interaktionssituation während der Begehung gut zu bestehen, dreht sich in der Vorbereitung auf die RAE sehr viel um die Optimierung von Outputgrößen.

Jeder Vorbereitungsprozess umfasst eine Vielzahl von (Selektions-)entscheidungen und die Institute gehen in ihren Entscheidungsfindungen sehr unterschiedlich vor (vgl. auch Gülker/Matthies/Matthäus 2009). Für Evaluationen der Leibniz-Gemeinschaft und im Rahmen des SEP gilt es, diejenigen Forschungsschwerpunkte und -projekte zu definieren, die in der Präsentation besonders betont werden sollen – angesichts der begrenzten Aufnahmekapazität der Gutachter/innen innerhalb der begrenzten Zeit lässt sich in aller Regel nicht die vollständige Vielfalt an Forschungsthemen darstellen. Dieser Selektionsprozess hat bereits einen Auftakt im Rahmen der Erstellung der vorbereitenden Unterlagen, auch hier werden bereits Schwerpunkte genannt und Zukunftsstrategien entwickelt. Bei der Begehung wird zusätzlich fokussiert und auf Postern sowie in mündlichen Beiträgen werden diejenigen Beiträge hervorgehoben, von denen man annimmt, dass sie auf die Bewertung einen positiven Einfluss haben können.

In der Regel schaffen die Institute für diese Zeit der Evaluationsvorbereitung eine eigene befristete Struktur. Ein kleines Steuerungsgremium organisiert den Prozess und je nach Partizipationsanspruch werden Leitungsebenen oder auch alle Mitarbeiter/innen in unterschiedlichen Phasen unterschiedlich intensiv in die Vorbereitungen einbezogen. Nicht selten werden auch grundlegende Umstrukturierungen des Instituts in Vorbereitung auf eine Evaluation vorgenommen:

> *„Ja und das [die bisherige Struktur, Anm. d. Autorinnen] war dann irgendwann natürlich auch irgendwie mal verbraucht, man brauchte also auch ne neue kreative Idee und die wollten wir den Gutachtern auch vorstellen. Und dazu gab es im Jahr davor dann natürlich einen Workshop, wenn man so will, wo wir wirklich raus sind aus dem Haus und uns für zwei Tage eingemietet haben am Rande von Berlin, um in Ruhe arbeiten zu können*

und haben uns die neue Struktur überlegt, ja wo hängt es, was müsste man besser machen. Und da sind dann die Forschungsfelder, so haben wir das dann genannt, rausgekommen als Vorschlag für die Gutachter dann auch, und zwar um das Miteinander-Arbeiten der verschiedenen Wissenschaftsrichtungen [zu zeigen]." (Illing: 76-86)

In diesem Beispiel werden zwei unterstellte Erwartungen der Gutachter/innen antizipiert, die sich häufig in den Schilderungen finden: Ein Institut sollte „in Bewegung" sein – veraltete Strukturen müssen überdacht und neu justiert werden. Und in einem Institut sollte interdisziplinär zusammengearbeitet werden. Eine weitere Anforderung, die in den schriftlichen Unterlagen zur Evaluation betont und auch als Kriterium bei den Gutachter/innen/n angenommen wird, ist die „Konsistenz" des Instituts. Entsprechend wird bei der inhaltlichen Darstellung wie bei der Organisationsstruktur darauf geachtet, dass ein übergreifendes „Profil" zu erkennen ist und dass das Spezielle oder auch das „Alleinstellungsmerkmal" deutlich wird. Konsistenz und Gemeinsamkeit gilt es auch im Auftritt zu demonstrieren und der Auftritt wird gründlich geübt:

„Und auch die Art und Weise, die Sprachregelung, an der arbeiten wir immer noch, also das haben wir gemerkt bei der ersten Probe, dass es wichtig ist, dass das Direktorium mit einer Zunge spricht, nicht mit den gleichen Worten, aber mit der gleichen Aussage. Und da war's uns aber vielleicht noch gar nicht so bewusst, so richtig bewusst geworden ist es uns erst bei der zweiten Probeevaluierung. Wir haben nämlich insgesamt drei Probeevaluierungen gehabt." (Dittmer: 279-285)

Auch wenn bei der Präsentation im Rahmen der Begehung das Institut als Ganzes im Vordergrund steht, lösen sich individuelle Interessen darüber allerdings nicht auf:

„Ach das sind verschiedene Mitarbeiter, also das sind vor allen Dingen so die jüngeren, also oder neuen Mitarbeiter, die sich noch nicht wiedergefunden haben dort, das ist also unter den alten Mitarbeitern eigentlich wenig, aber da kommen dann auch denke ich mal so Sachen da auf – also wir präsentieren uns ja nicht nur über die Forschung, sondern eigentlich über die Personen, die diese Forschung betreiben. Und wenn jetzt die Leute aber meinen, dass sie sich nicht wirklich wiederfinden oder sich nicht wirklich darstellen können, ist ja auch für ihre Karriere ein wichtiger Punkt, sich also zu platzieren und da also hier bestimmte Stellen zu besetzen und sagen, da bin ich und wenn jemand in dem Bereich was wissen will, dann sprecht bitte mich an. Ja, das wurde wenig, auf diesen Aspekt, der war eigentlich auch nicht wichtig, aber der wird in letzter Zeit also wichtiger." (Jenicke: 144-155)

Tatsächlich handelt es sich bei der Vorbereitung auf eine Begehung auch um einen komplexen Aushandlungsprozess, der für die Identifikation der Mitarbeiter/innen mit dem Institut wie für das kollegiale Miteinander positive wie negative Konsequenzen haben kann.

Im Rahmen der Vorbereitung auf die RAE werden ebenfalls zahlreiche (Selektions-)entscheidungen getroffen und auch hier werden die Entscheidungsprozesse mal mehr und mal weniger partizipativ organisiert. Ein Großteil der Entscheidungen bezieht sich hier allerdings auf das unmittelbare strategische Kalkül im Zusammenhang mit der Budgetformel, in die die Einzelbewertungen eingehen. Etwa wird angenommen, dass es bereits einen Unterschied machen kann, bei welcher Gutachtergruppe die Bewerbung eingereicht wird – in manchen Subdisziplinen kommen mehrere in Frage und manche gelten als strenger als andere.

Weiterhin kann als eine Art grundlegende strategische Entscheidung angesehen werden, wie viele Mitarbeiter/innen man als „forschungsaktiv" meldet. Dahinter stehen taktische Überlegungen: Geht man davon aus, dass die Budgetformel nicht linear ist, dass also sehr gute Gesamtnoten ein Mehrfaches an Budget versprechen, kann es rational sein, eher weniger und dafür nur die als Spitzenforscher/innen angenommenen Mitarbeiter/innen zu melden. In dem Zusammenhang ist es auch durchaus üblich, zusätzliche publikationsstarke Wissenschafter/innen von anderen Universitäten und Forschungseinrichtungen zu rekrutieren. Seitens des RAE-Managements wurde auf diese Taktik bereits im Laufe der Verfahren reagiert und festgelegt, dass alle gemeldeten Wissenschaftler/innen eine bestimmte Mindestvertragslaufzeit und ein Mindestvolumen am jeweiligen Institut nachweisen müssen.

„Game playing" ist ein viel genutzter Ausdruck im Zusammenhang mit der Vorbereitung auf die RAE. Und ein Maß an stragetischem Kalkül erscheint nötig, taktisch unkluges Verhalten (beispielsweise die Benennung zu vieler Forscher/innen) aus dem vergangenen Verfahren wird korrigiert. Allerdings finden alle taktischen Überlegungen unter Unsicherheit statt, tatsächlich wird der genaue Verteilungsschlüssel des Budgets vorab nicht kommuniziert, man weiß nicht, welche Note wie viel wert ist. Ein Interviewpartner kommt vor diesem Hintergrund zu einer recht fatalistischen Gesamteinschätzung:

> „*I see the process as something a kin to lottery and to gambling, and I'm an actually an individual who I can say at this point in my life which you know, I have never gambled, I have never put any money on a horse, or a dog, or on the lottery, I have never bought a national, I never bought a lottery ticket, no no, I have never, I never do things like that in terms of my temperament, and my personality, and yet when it comes to the RAE there is lots of gambles, because there's lots of points at which you say, could be this could be that, could do that could do this, and ultimately you have to decide, do I put this paper in or this papers in, and every, you know while some decisions are clear cut others are more borderline, including should we put individual A in or should we not on the borderline, and so on." (Gomati: 785-796)*

Neben diesen taktischen Überlegungen steht aber die Auswahl geeigneter Publikationen im Zentrum der Vorbereitung auf die RAE. Dieser Auswahlprozess beginnt vielfach mehrere Jahre vorher, häufig wird institutsintern eine RAE simuliert, die Bewertung der Papiere wird mal vom Institutsdirektor allein, mal von einem internen Gremium und mal unter Einbezug externer Gutachter vorgenommen. In dem Maße, wie die RAE auf individuelle Leistungen konzentriert ist, sind auch hier die Aushandlungsprozesse in der Vorbereitung entsprechend komplex und interessengeleitet. Nicht allein für die Forschungseinheit, sondern auch für die einzelnen Mitarbeiter/innen hängt vom Ergebnis der RAE viel ab:

> *„We try to make it as clear as possible that a university has to make decisions in what it perceives to be its best interest and that there are no necessary implications for people, for an individual's career progression, through inclusion or exclusion of the RAE. However, in the real world, it is damaging for an individual career, I think, not to be included. Who is and who isn't included becomes public, after the results of the RAE. Individuals aren't graded, so you can't say, "Bob's got five stars and Jim's got three stars." But we will be able to see who was an who wasn't included. So I guess, potentially, if people are applying for posts beyond the university level, potential employers can look at the RAE returns in 18 months' time, and see that Bob wasn't included in the INSTITUT's RAE return. So in the real world, you do want to be in, as an individual, and it is a source of anxiety for people, if they're not included." (Uhtke 270-281)*

Evaluationen sind also unabhängig vom Verfahrenstyp einschneidende Ereignisse im Alltag eines Forschungsinstituts wie in individuellen Karrieren. Die Frage ist, ob und inwiefern diese Ereignisse auch nachhaltige Auswirkungen auf den wissenschaftlichen Alltag haben. Einige Indizien dazu werden im Folgenden dargelegt.

4 (Wie) verändert sich wissenschaftlicher Alltag?

Wenn wir im Folgenden nach Veränderungen im „wissenschaftlichen Alltag" fragen, dann beziehen wir uns auf die Ebene der Organisation von Wissenschaft und auch hier stützen wir uns nicht auf eigene Beobachtung, sondern rekonstruieren Deutungen der Beteiligten. Über die „epistemische Praxis" kann im engen Sinn keine Aussage gemacht werden.

Wie bereits dargestellt, nehmen Institute in der Vorbereitung auf Evaluationen zum Teil erhebliche Umstrukturierungen vor. Dies gilt sowohl für die deutschen und niederländischen Institute als auch für die in Großbritannien. Dies ist bemerkenswert, weil die RAE keinerlei inhaltliches Feedback vorsieht. Die Institute erhalten als Ergebnis des Verfahrens lediglich eine Zahl, beziehungsweise mit dem 2008er Verfahren erstmals ein Zahlenprofil, das heißt eine Tabelle, in der zu erkennen ist,

wie viele der eingereichten Arbeiten mit welcher Note versehen wurden. Und dieses Ergebnis löst in den Instituten eine enorme Aktivität aus. Welcher Art diese Aktivität ist, hängt dann allerdings ganz davon ab, welche Organisationsphilosophie, welche Leitungsvorstellung oder eben welche Idee darüber besteht, wie wissenschaftliche Bestleistung organisatorisch am besten sicher zu stellen ist. Dies wird im Folgenden an zwei Beispielen illustriert.

Institut A ist ein hoch renommiertes Institut in seinem Fach und hat bei der vorherigen RAE nicht die beste Note erhalten. In Reaktion darauf wurde unmittelbar noch im Jahr 2002, als die Resultate veröffentlicht wurden, ein neuer Direktor eingestellt. Dieser hat dann umfassende Umstrukturierungsmaßnahmen durchgeführt, zusammenfassend:

> „*So the first thing was to shift the balance between teaching and research, second thing was the research monitoring scheme, the third thing was support and the fourth thing really was to try and organise ourselves better, as a research organisation, no as a research institute. That we'd really been a bunch of a hundred individuals who do our own thing. And we weren't used to working together very much, so one of the things, the fourth thing that I pushed, was organising ourselves in some cases into research centres, where there really was a kind of coherent group of people interested in the same sort of thing so they might then have workshops or conferences and additional seminars and so on, so it would become not just an individual activity but a collective activity." (Jäger: 99-110)*

Die Kernidee der Umstrukturierungen war also ein neuer organisatorischer Rahmen, der inhaltliche Zusammenarbeit besser ermöglichen sollte. Innerhalb dieser neu geschaffenen kleinen Fachgruppen wurde zudem ein „research monitoring scheme" eingeführt: Für jede/n Wissenschaftler/in wurden zwei Kollegen/innen benannt, mit denen ein Fünfjahresplan entwickelt und jährlich überprüft wurde. Bei Nichteinhalten wurden, auch flankiert durch zusätzliche finanzielle Mittel, Entlastungsstrategien gesucht, die intensivere Forschung ermöglichen sollten. Die Fachteams waren dann auch zuständig dafür, Vorschläge für die einzureichenden Artikel zu machen. Als organisatorisches Konzept steht also kollegiale Zusammenarbeit hier im Vordergrund – der fachliche Rat, aber auch die soziale Kontrolle unter Kollegen wurden für geeignet gehalten, wissenschaftliche Leistung zu verbessern.

Im *Institut B* hat sich der Direktor vorgenommen, mit dem Ergebnis der RAE die Sichtbarkeit der guten Forschung noch einmal deutlich zu erhöhen. Das Ergebnis des vorherigen Verfahrens war entsprechend der Kapazität des Instituts zwar zufriedenstellend, aus Sicht des (2004 neu ins Amt gekommenen) Direktors aber steigerbar. Mit diesem Ziel hat er noch in 2004 intern eine RAE simuliert. Alle Unterinstitute waren aufgefordert, Bewerbungen zu formulieren und Artikel zu identifizieren, die bewertet werden sollten. Hintergrund dieses Verfahrens war die Überlegung, dass in der kommenden RAE nicht (wie in der letzten) möglichst viele, sondern tatsächlich nur die besten Artikel eingereicht werden sollten. Frühzeitig sollten diejenigen Wissenschaftler/innen identifiziert werden, deren Arbeiten berücksichtigt werden sollten und damit auch diejenigen, deren Arbeit nicht mit eingereicht werden würde:

> „*I was more concerned with the overall quality and if it meant that we were reducing the number of people that we were going to submit then that was fine, because it demonstrated that we were really pushing for quality. So in 2005 we decided that we would, then, we decided that some of our colleagues would not be submitted. They were then asked to do other things like do more teaching or do more administration to free up the more active researchers to do a lot more over the next 2 or 3 years." (Fredo: 86-92)*

Anders als im Falle des Instituts A wird hier nicht angenommen, dass durch intensivere Zusammenarbeit der gemeinsame Output erhöht werden kann, sondern das organisatorische Konzept stellt den individuellen Wettbewerb zwischen den Wissenschaftler/inne/n ins Zentrum. In dem Maße, wie um Zeit und Ressourcen konkurriert wird, ist es rational (nur) in diejenigen zu investieren, von denen zum Zeitpunkt X auch ein positives Resultat erwartet wird – die als

weniger leistungsstark eingeschätzten Mitarbeiter/innen werden zu Gunsten der leistungsstark eingeschätzten Mitarbeiter/innen zum Verzicht auf Forschungszeit angehalten. Diese Unterteilung der Belegschaft hat dabei durchaus über die aktuelle RAE hinaus Bedeutung:

„*On advice from HR [Human Resources], we also came to the agreement that it didn't necessarily mean their research career was at an end, what it meant was after the RAE, that from about, last year, September 2007 that the people who were not going to be submitted would be invited to submit a research plan to their head of school, and the research plan would be discussed with their head of school. If the research plan was accepted, then we would support those people to become research active again. If the research plan was not accepted, then they would be asked to continue with a double load of teaching, more administration, they wouldn't be research active. And then there would be a change in the terms and conditions." (Fredo: 171-180)*

Mit den Ergebnissen der RAE sind also tiefgreifende Veränderungen für eine Forschungsorganisation verbunden und die Beispiele legen nahe, dass sich an diesen beiden Instituten auch die Koordination, Kooperation und auch Kommunikation zwischen den Wissenschaftler/inne/n verändert hat – an Institut A findet fortan ein bislang nicht geführter Austausch mit Kolleg/inn/en statt, an Institut B werden sich alle mit großer Energie auf die individuellen Leistungen konzentrieren, um nicht Gefahr zu laufen, bei der nächsten RAE „aussortiert" zu werden. In welchem Falle damit eine „leistungsfähigere" Wissenschaft verbunden ist, kann nicht eingeschätzt werden, aber es wird deutlich, wie das gleiche Ziel, nämlich gut bzw. besser abschneiden zu wollen als beim letzten Mal, zu völlig unterschiedlichen Strategien führen kann.

Für die Evaluationen der Leibniz-Gemeinschaft und nach dem SEP würde man dagegen annehmen, dass hier die Ergebnisse mit eindeutigeren strategischen Orientierungen einhergehen. Anders als im Falle der RAE sind hier im Ergebnisbericht ja ausführliche inhaltliche Rückmeldungen und auch Handlungsempfehlungen für die weitere Organisationsentwicklung vorgesehen. Und tatsächlich wird sich mit den Empfehlungen sehr intensiv auseinandergesetzt, dies auch, weil in der nächsten Evaluation danach gefragt wird, wie damit umgegangen wurde. Allerdings treffen – wie bei der RAE – auch hier die Ergebnisse auf einen bereits vorhandenen Rahmen, auf ein Konzept dazu, wie wissenschaftliche Leistung organisatorisch am besten zu gewährleisten ist (vgl. Simon 2008). Und so werden von den Empfehlungen diejenigen umgesetzt, die intern anschlussfähig sind:

„[…] vielleicht zwei Drittel, ich sage mal ein Drittel der Empfehlungen haben wir mehr oder weniger direkt umgesetzt […] ein Drittel empfanden wir nicht wirklich als relevant und ein Drittel der Empfehlungen ist ungefähr so, dass es unrealistisch ist sie umzusetzen". (Nogat: 11 – 15)

Evaluationen begründen damit in der Regel keine grundlegend neuen Institutsstrukturen, aber sie werden als Validierungs- und Legitimationsinstanz für laufende interne Reformen genutzt (vgl. Meier/Schimank 2010).

In einem niederländischen Beispiel etwa wurde ein Institut komplett neu strukturiert – der alte Direktor schied einige Jahre vor der Evaluation aus, danach folgten mehrere kommissarische Leitungen, bevor mit dem aktuellen Direktor dann ein grundlegender Neuanfang gemacht wurde. In dem Zusammenhang wurde auch eine große neue Abteilung eingerichtet, die dem gesamten Institut mittelfristig ein neues Profil geben würde. Die Evaluation wird vorab auf diesem Weg als wichtige Validierungsinstanz wahrgenommen:

„*So, in that way, in that sense the new evaluation in 2007 came at a good point because well, we have a lot of plans and things and we are putting well, new research groups and so it was interesting, I was, it was an interesting thing to hear, what the committee thought about that, so it was, that really function and well I was anxious to hear what they thought about it." (Yukawa: 67-72)*

Im Ergebnis wurde dieses Institut dann insgesamt sehr gut bewertet, Kritik gab es allerdings

an der Qualität (nicht an der inhaltlichen Ausrichtung) der neu eingerichteten Abteilung. In der Reaktion auf den Bericht und in der internen Reflexion wurde dies darauf zurückgeführt, dass die Abteilung noch am Anfang stehe und es in der Präsentation nicht gut gelungen sei, die eigentliche Qualität der wissenschaftlichen Arbeit dort herauszustellen. In der Gesamtstrategie sah sich der Direktor demnach durch die Evaluation bestärkt und im Hinblick auf die konkrete Abteilung zwar herausgefordert, aber nicht prinzipiell in Frage gestellt.

In einem deutschen Fall ist die Validierung einer neuen Struktur durch die Evaluation im Prinzip nicht gelungen, was aber auch hier nicht zu grundlegenden Erschütterungen geführt hat. Das Institut hatte zwei Jahre vor dem Evaluationstermin eine neue matrixartige Struktur eingeführt und nun die Präsentation auch an den neuen Forschungsschwerpunkten orientiert. In der Darstellung wurde diese neue Struktur von den Gutachter/innen aber wohl nicht problemlos verstanden und auch im Bewertungsbericht kritisiert. In der Reflexion der Institutsvertreter/innen werden hier Schlussfolgerungen wiederum lediglich in Bezug auf die Präsentationstechnik und nicht in Bezug auf die Organisationsstruktur selbst gezogen. Oder auch das Verständnisvermögen der Gutachter/innen wird in Zweifel gezogen:

> *„Vielleicht bringen die es auch von Haus aus mit viele und dann gibt es eben so eine hierarchische Struktur nur und das versucht man wiederzufinden. Und dieses Matrixschema, was hier existiert, das ist auch gar nicht so leicht zu verstehen, dass also sagen wir mal auf dem kleinen Dienstweg von A nach B gegangen werden kann, ohne dass das hier irgendwie große Wellen schlägt." […]*
>
> *Und das kann natürlich auch kein anderes Institut bieten, weil wie gesagt das gibt's eben nicht noch mal und deswegen ist es natürlich auch schwer verständlich, jemand, der ja eben normalerweise sein Thema hat und das bearbeitet und rechts und links ist da auch nichts mehr, jedenfalls in dem Institut nicht, der kann nur schwer verstehen, dass es jetzt von A bis Z durchgeht" (Xaver-Unger: 786-791; 842-847)*

Dies also als Beispiele dafür, wie Empfehlungen dann nicht inhaltlich aufgenommen werden, wenn sie an vorhandene Konzepte des Instituts nicht anschlussfähig sind – der Umgang damit ist dann strategischer Art, die Darstellung soll beim nächsten Mal verbessert werden (vgl. Leisyte/Enders/de Boer/Harry 2010). Ähnliche Mechanismen finden wir auch in Bezug auf inhaltliche Schwerpunktsetzungen, die im Rahmen der Empfehlungen vorgeschlagen werden. So wird etwa in einem Institut vorgeschlagen, die Drittmittel zu steigern, mehr zu publizieren und den Anteil an Politikberatung zu erhöhen. Alle drei (zueinander auch in potenziellen Zielkonflikten stehenden) Empfehlungen werden reflektiert und erheblich relativiert. Abgewogen wird etwa der Ressourcenaufwand für zusätzliche Drittmittel und für die Politikberatung wird festgestellt, dass sie gerade bei jüngeren Kolleg/inn/en auch kontraproduktiv für die wissenschaftliche Karriere sein kann.

Inwiefern sich der wissenschaftliche Alltag in Folge von Evaluationen der Leibniz-Gemeinschaft oder nach dem SEP verändert ist entsprechend schwierig nachzuvollziehen. Erkennbar ist, dass der Evaluationstermin zur internen Reflexion zwingt und damit auch Reformschritte terminiert werden, die ansonsten möglicherweise verschoben oder auch ganz aus den Augen geraten würden. Und in dem Maße, wie institutionelle Reformen zu einem Wert an sich werden und Begriffe wie „Profilschärfung" als Orientierung dienen, ist vorstellbar, wie der individuelle Forschungsbeitrag zu diesem Gesamtprofil nachgefragt und erklärungsbedürftig ist. Wiederum können die Strategien zur Motivation der einzelnen Mitarbeiter/innen ganz unterschiedlich sein.

5 Fazit

Mit den Evaluationen der deutschen Leibniz-Gemeinschaft und dem niederländischen SEP auf der einen Seite und der RAE in Großbritannien auf der anderen Seite wurden hier zwei unterschiedliche Typen von Institutsevaluationen vorgestellt und auf ihre institutionellen Effekte hin überprüft. Während bei der Leibniz-Gemeinschaft und beim SEP die Begehung der Einrichtung und damit die Kommunikation zwischen Evaluatoren und Evaluierten ein zentraler Verfahrensschritt ist und sich die Bewertung auf die Gesamtleistung einer Organisation bezieht, werden in der RAE in erster Linie individuelle Forschungsleistungen bewertet und dann für die Einrichtung aggregiert.

Die institutionellen Effekte der beiden Verfahren sind ähnlicher, als man zunächst vielleicht annehmen würde: Bei den Evaluationsverfahren der Leibniz-Gemeinschaft und der des SEP sind sie eine im Verfahren vorgesehene Folge, aber auch in der RAE, die auf individuelle Erfolge von Wissenschaftler/innen abzielt, finden Veränderungen in der Organisationseinheit statt. Sicherlich ist erkennbar, wie die RAE einen größeren Druck auf alle Institutsmitarbeiter/innen inklusive Leitungsebene ausübt. Die unmittelbare Verknüpfung von Ergebnis und Budgethöhe wiegt dabei ähnlich schwer wie der Reputationsgewinn oder -verlust, der institutionell wie individuell mit diesem bereits zur Tradition gewordenen Verfahren verbunden ist.

Als zentraler Effekt für beide Verfahren lässt sich aber eine Stärkung der Organisationsebene in der Wissenschaft feststellen, deren Auswirkungen auf den wissenschaftlichen Alltag bedeutsam sein dürften, deren Wirkrichtung wir aber bislang nicht beobachten können. Institutsleitungen werden durch Evaluationen gestärkt (vgl. Meier/Schimank 2010). Im Falle der RAE kann es etwa von der Bewertung der Institutsleitungen abhängen, ob Wissenschaftler/innen weiter forschungsaktiv sein können oder nicht; Monitoringinstrumente werden eingeführt, wie wir sie sonst aus unternehmerischen Zusammenhängen kennen. Im Falle von Leibniz-Gemeinschaft und SEP sind Institute gefordert, Aktivität in Richtung gemeinsamer Forschungsstrategie und Profil zu demonstrieren – Leitungen nutzen die Evaluationen zur Legitimationsbeschaffung für interne Struktur- und Organisationsreformen. Evaluationen haben also Folgewirkungen in den Forschungseinrichtungen, die den wissenschaftlichen Alltag maßgeblich tangieren und über eine rein rhetorische Bedienung von antizipierten Vorstellungen der Evaluatoren über „gute" Wissenschaft hinausgehen.

Die Konzepte dazu, wie wissenschaftlich erfolgreiche Arbeit organisatorisch unterstützt werden kann, sind dabei auch innerhalb der nationalen Forschungstraditionen höchst unterschiedlich. Hier weisen die Befunde auf wesentlichen Forschungsbedarf hin: Die organisationssoziologische Überprüfung dieser vielfach impliziten (aber in der Praxis einschneidenden) Konzepte steht weiterhin aus.

Literatur

Boni, Manfred, 2010: Analoges Geld für digitale Zeiten: der Publikationsmarkt der Wissenschaft. Leviathan 3/2010, 293-312.

Frey, Bruno, 2008: Evaluitis – eine neue Krankheit, in: *Matthies, Hildegard / Simon, Dagmar (Hg.):* Wissenschaft unter Beobachtung – Effekte und Defekte von Evaluationen. Wiesbaden: VS Verlag, 125-140.

Gülker, Silke / Matthies, Hildegard / Matthäus, Sandra, 2009: Evaluationsverfahren aus labor-konstruktivistischer Perspektive. WZB Discussion Paper SP III 2009-601. Berlin: Wissenschaftszentrum Berlin für Sozialforschung. Online: http://bibliothek.wzb.eu/pdf/2009/iii09-601.pdf.

HEFCE, Higher Education Funding Council, 2005: RAE 2008.Guidance on submissions. Online: http://www.rae.ac.uk/pubs/2005/03/rae0305.doc (6/2008).

Heintz, Bettina, 2006: Governance by Numbers. Zum Zusammenhang von Quantifizierung und Globalisierung am Beispiel der Hochschulpolitik, in: *Schuppert, Folke / Voßkuhle, Andreas (Hg.):* Governance von und durch Wissen. Baden-Baden: Nomos Verlag, 110-128.

Kieser, Alfred, 2010: Unternehmen Wissenschaft? Leviathan, Heft 3/2010, 347-369.

KNAW / VSNU / NOW, 2003: Standard Evaluation Protocol 2003-2009 for Public Research Organisations. Amsterdam; Utrecht; Den Haag.

Leisyte, Liudvika / Enders, Jürgen / de Boer, Harry, 2010: Mediating Problem Choice – Academic Researchers' Responses to Changes in their Institutional Enviroment, in: *Whitley, Richard / Gläser, Jochen / Engwall, Lars (Hg.):* Reconfiguring Knowledge Production. Changing Authority Relationships in the Sciences and their Consequences for Intellectual Innovation. Oxford: Oxford University Press, 266-290.

Martin, Ben / Whitley, Richard, 2010: The UK Research Exercise: Al Case of Regulatory Capture?, in: *Whitley, Richard / Gläser, Jochen / Engwall, Lars (Hg.):* Reconfiguring Knowledge Production. Changing Authority Relationships in the Sciences and their Consequences for Intellectual Innovation. Oxford: Oxford University Press, 51-80.

Meier, Frank / Schimank, Uwe, 2010: Mission Now Possible: Profile Building and Leadership in German Universities, in: *Whitley, Richard / Gläser, Jochen / Engwall, Lars (Hg.):* Reconfiguring Knowledge Production. Changing Authority Relationships in the Sciences and their Consequences for Intellectual Innovation. Oxford: Oxford University Press, 211-238.

Simon, Dagmar, 2008: Als Konsequenz mehr Kohärenz? Intendierte und nicht intendierte Wirkungen von institutionellen Evaluationen, in: *Matthies, Hildegard / Simon, Dagmar (Hg.):* Wissenschaft unter Beobachtung – Effekte und Defekte von Evaluationen. Wiesbaden: VS Verlag, 161-177.

Georg Rudinger, Norbert Hilger[1]

Ausstieg aus dem CHE-Ranking

Ausgangslage

Im September 2009 fasste das Rektorat der Rheinischen Friedrich-Wilhelms Universität Bonn den Beschluss, bis auf weiteres am CHE-Ranking nicht mehr teilzunehmen – die Gründe hierfür werden im Folgenden dargelegt. Daraufhin fand im Januar 2010 ein Gespräch mit Vertretern des Centrum für Hochschulentwicklung (CHE), u.a. mit dem Leiter Prof. Dr. Ziegele, statt, in dem die Bonner Bedenken erörtert wurden. Diese bezogen sich vor allem auf die mangelnde Wissenschaftsorientierung des Fragenkatalogs, auf die Eignung bzw. Nichteignung von (quantitativen) Indikatoren für die Messung von Qualität der Wissenschaft, auf die mangelnde Differenzierung zwischen den Fächern, vor allem Fächerkulturen, und auf die mangelnde Transparenz. Insbesondere sind damit Fragen zur Methodik, zu Rücklaufquoten und zur Repräsentativität gemeint und damit verbunden zum (seinerzeitigen) „Ampel"-Ranking, bezüglich der Farbsymbolik zum einen und vor allem zu den Rang-Gruppen auf der Basis von sogenannten Fehlerbalken zum anderen. Im Mai 2010 erhielt die Universität Bonn ein schriftliches Feedback von Prof. Ziegele, welches im Folgenden ebenso mitberücksichtigt wird wie die Präsentation von Prof. Ziegele am 31. Januar 2011 vor der Landesrektorenkonferenz der Universitäten in Nordrhein-Westfalen (LRK), wo auf diese und auch von anderer Seite erhobene Bedenken eingegangen wurde; berücksichtigt wurden ebenfalls die Ergebnisse weiterer Gespräche vor (November 2010) und nach (Mai 2011) dieser Präsentation. Es handelt sich also um einen Werkstattbericht, für dessen Erstellung der Vortrag auf der iFQ-Jahrestagung im Juni 2010 eine wesentliche Rolle spielte, der aber über den Vortrag hinausgeht.

1 Kennzeichen des CHE-Rankings gemäß Selbstbeschreibung des CHE[2]

Das Ziel des Rankings ist es, Studienanfänger und Hochschulwechsler über Studienmöglichkeiten und -bedingungen zu informieren und die Angebots- und Leistungstransparenz im Hochschulbereich zu verbessern. Zu diesem Zwecke werden Fakten und Urteile unter unterschiedlichen Perspektiven erhoben und zu Indikatoren verarbeitet[3]. Als Faktenbasis dient die Analyse objektiver Daten der Studiensituation. Urteile werden durch Einschätzung der Studierenden und durch Befragung der Professorinnen und Professoren in die Gesamtbewertung eingebracht. Diese Datenbasis wird fachbezogen und nicht hochschulweit aggregiert. Es geht also um einen Vergleich der Fächer und nicht um ein Gesamtranking von Hochschulen. Das Ranking – so das CHE – ist multidimensional, das heißt unter verschiedenen Perspektiven werden Indikatoren für die Bewertung erhoben, wie zum Beispiel die Publikationen, die Promotionen, Ausstattung (Bibliothek, Computerpool u. ä.) oder Drittmitteleinwerbungen. Das Ranking ist dadurch charakterisiert, dass nicht einzelne Rangplätze vergeben, sondern Rang-Gruppen gebildet werden: eine Spitzengruppe, eine Schlussgruppe und eine Mittelgruppe.

1 Rheinische Friedrich-Wilhelms-Universität Bonn, Zentrum für Evaluation und Methoden – ZEM. Die Autoren danken Frau Dipl.-Psych. Katharina Olejniczak für ihre Unterstützung.

2 http://www.che-ranking.de/methodenwiki/index.php/Hauptseite

3 http://www.che-ranking.de/methodenwiki/index.php/Indikatoren

2 Erhebung von Fakten[4]: zum Beispiel bibliometrische Analyse[5]

Die im Folgenden näher beleuchteten bibliometrischen Analysen sind Bestandteil des CHE-ForschungsRankings[6]. Das ForschungsRanking enthält jeweils fachspezifisch Informationen zu den Indikatoren „Drittmittelausgaben", „Publikationen", „Erfindungen", „Promotionen" und „Reputation". Der Indikator Reputation wird allerdings nicht zur Identifizierung der Gruppe „forschungsstarke Hochschulen je Fach" herangezogen, sondern lediglich als zusätzliche Information.

Absicht ist es von Seiten des CHE, Aktivitätsindikatoren zu ermitteln, welche die Teilnahme an fachwissenschaftlicher Forschungskommunikation indizieren, darüber hinaus die Bewertung von Forschungsleistungen über die Publikationen zu erhalten und letztendlich die Resonanz in der Fachöffentlichkeit als Indikator für Qualität der Forschung festzustellen. Die Kritik, welche von der Bonner Seite erhoben worden ist, umfasst folgende Punkte: Die fachbereichsbezogene bibliometrische Analyse orientiert sich an der Quantität, d.h. an der Länge der Publikationen, wird dann je nach Fach gewichtet und entsprechend gepunktet. Das bedeutet nach Bonner Meinung eine Verdoppelung der Quantifizierung. Zu einem Teil geschieht diese Quantifizierung, ohne eine qualitative Zeitschriftenklassifikation zu berücksichtigen. Der Bonner Meinung nach bilden die resultierenden Ergebnisse somit nicht zugrundeliegende Forschungsleistung ab. Die schon erwähnten verschiedenen Fachkulturen sind nicht berücksichtigt, weil über Qualität von Monographien keinerlei Aussage gemacht wird[7].

Der CHE-Standpunkt dazu ist der Folgende: Die einzubeziehenden Publikationen werden durch Abfragen in den Literaturdatenbanken ermittelt, die auf den Vorschlägen der Fachbeiratsmitglieder basieren. Dieser Auszug sollte zudem die wesentlichen Publikationstypen des Faches abdecken. Analysen in den Datenbanken von Thompson Scientific werden durch das Forschungszentrum Jülich durchgeführt. Für Fächer, bei denen die Verwendung dieser Datenbank unzureichend erscheint, werden zumeist nationalorientierte Datenbanken herangezogen. Die Eignung der Datenbanken wird in jedem Zyklus neu diskutiert und falls nötig geprüft. So stehe man z. B. mit der GESIS in Verhandlungen bezüglich des Einbezugs von SOLIS - Social Science Literature Information System. Man geht also in (fachspezifische) Datenbanken. Unter www.che-ranking.de/methodenwiki/index.php/Bibliometrische_Analyse finden sich tabellarisch Datenbanken und Spezifika nach Fächern, von denen hier nur einige Beispiele wiedergegeben werden:

- **Bibliometrische Analyse Anglistik**
  - Datenbasis für die Publikationsanalyse: Annual Report on English and American Studies (AREAS).
- **Bibliometrische Analyse Mathematik, Naturwissenschaften, Medizin**
  - Datenbasis: Zitationsdatenbanken SCI/SSCI und A&HCI des Web of Science (WoS) (durchgeführt vom FZ Jülich - Zentralbibliothek).
- **Bibliometrische Analyse Wirtschaftswissenschaften**
  - Eine Basis SCI/SSCI/A&HCI des ISI Web of Science, eine weitere die Datenbanken aus dem WisoNet, u.a. HWWA (Institut für Wirtschaftsforschung Hamburg), ECONIS (Institut für Weltwirtschaft Kiel) und BLISS (GBI München).

4 http://www.che-ranking.de/methodenwiki/index.php/Datenerhebungen#Erhebung_von_Fakten

5 http://www.che-ranking.de/methodenwiki/index.php/Bibliometrische_Analyse

6 http://www.che-ranking.de/methodenwiki/index.php/ForschungsRanking_Indikatoren

7 vgl. dazu Alfred Kieser, FAZ vom 09.06.2010 „Die Tonnenideologie der Forschung - Ranking, Rating, Bibliometrie". http://www.faz.net/artikel/C31373/akademische-rankings-die-tonnenideologie-der-forschung-30250841.html

- Für BWL und VWL wurde zudem ein Indikator ermittelt, der den Schwerpunkt auf internationale Sichtbarkeit legt. Die vom FZ Jülich durchgeführte Datenerhebung basiert auf der Abfrage des ISI Web of Science, speziell auf den Online-Versionen der Datenbanken Science Citation Index Expanded, Social Sciences Citation Index und Arts & Humanities Citation Index.

Zurzeit sieht man hier beim CHE keinen Korrekturbedarf.

3 Erhebung von Urteilen[8]

Die Erhebung von Urteilen findet zum einen in der Professorenbefragung, zum anderen in der Studierendenbefragung statt. Die Professorenbefragung soll Indikatoren zur Reputation des Faches im Sinne einer Studier-Empfehlung und zur Reputation der Forschung der entsprechenden Fächer an anderen Universitäten liefern[9].

Aus der Studierendenbefragung ergibt sich eine Vielzahl von Indikatoren via online-Fragebogen[10], an den sich allerdings eine erste inhaltliche Bonner Kritik knüpfte, dass nämlich äußere Bedingungen anstatt Forschungs- und Wissenschaftsorientierung im Mittelpunkt stünden, d.h. die Lehrtraditionen von Universitäten würden weniger berücksichtigt als die durch die Bologna-Reform in den Vordergrund gestellten Qualitätskriterien durchorganisierter Studiengänge.

3.1 Inhaltliche Kritik

Bei der erwähnten Präsentation vor der Landesrektorenkonferenz sind von Herrn Prof. Ziegele einige Fragen in Ergänzung der Studierendenbefragung zur Stärkung des Wissenschaftsbezuges vorgestellt worden (s. Abbildung 1 aus der erwähnten Ziegele-Präsentation vor der LRK NRW).

Abb 1: Auszug aus dem Onlinefragebogen der Studierendenbefragung des CHE

Quelle: Präsentation von Frank Ziegele auf der Landesrektorenkonferenz der Universitäten in NRW am 31.01.2011

8 http://www.che-ranking.de/methodenwiki/index.php/Datenerhebungen#Erhebung_von_Urteilen

9 Aber auch Professoren können irren. „Ich muss Ihnen ein Geständnis machen", beginnt Phil Baty, der stellvertretende Herausgeber des Britischen Hochschulmagazins Times Higher Education, sein Bekenntnis: Das Ranking, auf das seit 2004 viele Universitäten auch in Deutschland jeden Oktober gespannt warten, habe bislang schwerwiegende Mängel aufgewiesen; vgl. http://www.spiegel.de/unispiegel/studium/0,1518,699747,00.html

10 Mehr zu den Fragebögen unter: http://www.che-ranking.de/methodenwiki/index.php/Datenerhebungen#Frageb.C3.B6gen_des_CHE-HochschulRankings

Die items zum Wissenschaftsbezug werden von den Studierenden gut angenommen, so CHE.

3.2 Methodische Fragen

Die Antworten zu den Aspekten der Studiensituation – bewertet auf einer Skala von 1 bis 6 – werden großteils über eine Indexbildung verdichtet. Die Beziehungen zwischen den Items wurden mit Hilfe von Reliabilitäts-Analysen überprüft.[11] Daran knüpften sich folgende (Bonner) Fragen, die für die Bonner Entscheidung nicht von zentralem Interesse sind, allerdings der testtheoretischen und längsschnittlichen Prägung der Autoren geschuldet sind:

- Ist die Index-Bildung von dem Ergebnis der Reliabilitätsanalyse abhängig?
- Mit welcher Methode wird die Skalenanalyse vorgenommen?
- Wie verhält es sich mit den Trend- und Veränderungsaussagen, wenn die Skalen über die Zeit nicht invariant wären?

Laut CHE wird diese Konsistenz-Prüfung hier und da vorgenommen, aber nicht so regelhaft und systematisch (Reliabilitätsanalyse), wie es im alten Methodenwiki (noch) steht. Die Indexwerte der Studierendenurteile entsprechen dem (ungewichteten) arithmetischen Mittel der Einzelbewertungen. Dagegen ist unter pragmatischer Perspektive nichts einzuwenden – gegeben die Eindimensionalität; damit beißt die Katze sich aber in den Schwanz!

4 Statistische Ermittlung von Ranggruppen[12]

Zur Einteilung in die Ranggruppen Spitzen-, Mittel- und Schlussgruppe (mit der – inzwischen seit 2011 – farblichen Kennzeichnung grün, gelb und blau) werden für Fakten und Studierendenurteile zwei grundlegend verschiedene Verfahren eingesetzt.

Bei den Fakten (z.B. Drittmittel, wissenschaftliche Veröffentlichungen, Erfindungsmeldungen, Promo-tionen) wird die Gruppenbildung nach Quartilen dergestalt vorgenommen, dass die Werte der Größe nach geordnet und dann in drei Gruppen aufgeteilt werden: Die ersten 25 % (1. Quartil) sind die Spitzengruppe, das zweite und dritte Quartil sind die Mittelgruppe und die letzten 25 % (4. Quartil) die Schlussgruppe.

Bei den Urteilen geschieht die Gruppenbildung nach signifikanten Abweichungen vom Durchschnittsurteil im Fach, d.h. den jeweiligen Mittelwerten der Studienbereiche. Denn im Unterschied zu den Fakten sind die Studierendenurteile abhängig von den jeweils antwortenden Studierenden, sie sind einer gewissen Unsicherheit unterworfen. Wie gut sie jeweils dem „wahren" Urteilswert für einen Fachbereich entsprechen, hängt wesentlich von der Zahl der Antwortenden und der Bandbreite ihrer Bewertungen ab. Es werden die Fächer in ihrer relativen Position zum Durchschnittswert für den gesamten Studienbereich lokalisiert und nach folgender „Logik" gruppiert: Liegt der (Gesamt-) Mittelwert für den jeweiligen Studienbereich außerhalb des Intervalls eines Faches an einer Hochschule (HS), kommt das entsprechende Fach in die Extremgruppe, entweder rot oder grün[13] , ansonsten in die (gelbe) Mittelgruppe (vgl. die folgende Abbildung).

In dieser Abbildung ist die senkrechte Linie der Mittelwert der Mittelwerte, die waagerechten farbigen Linien sind die Mittelwerte der Universitäten in einem spezifischen Fach/Studiengang/Studienbereich mit ihrem jeweiligen 95-prozentigen Konfidenzintervall.

Liegt der Mittelwert der Mittelwerte also im Konfidenzintervall, kommt die Hochschule in die Mittel-

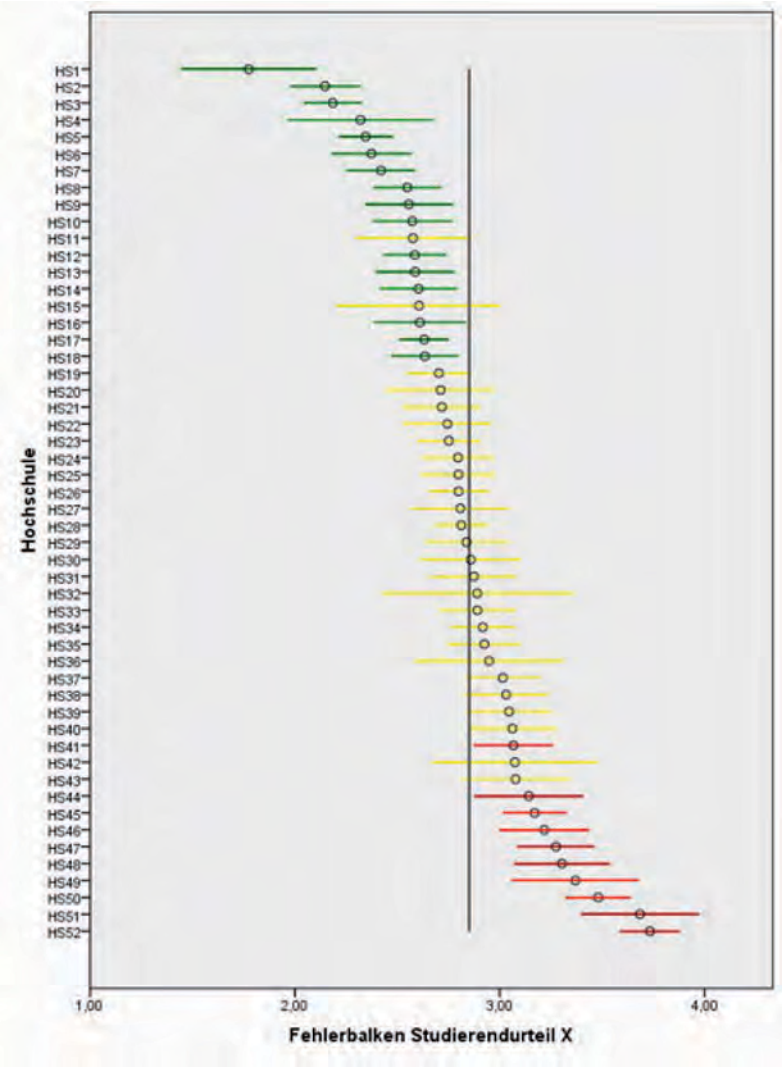
11 Arbeitspapier des CHE, Nr. 119 (Mai 2009), S.4/47.

12 http://www.che-ranking.de/methodenwiki/index.php/Statistische_Ermittlung_von_Ranggruppen

13 Dies waren die Ampel-Farben zur Zeit der Bonner Kritik; auf Veränderungen des Ansatzes und damit verbunden der Symbolik wird später eingegangen.

gruppe (gelb), liegt er außerhalb des Konfidenzintervalls in die Spitzen (grün)- oder Schlussgruppe (rot). Dies waren also – wie gesagt – bis zum letzten Jahr die der Ampelsymbolik geschuldeten Gruppenfarben im Ranking.

Abb. 2: Ranggruppenbildung Studierendenurteil



Quelle: www.che.de/methodenwiki

Diesen Definitionen und diesem Bild sind weitere kritische Detailfragen gefolgt, mit denen wir uns aber hier nicht aufhalten wollen, wie z.B.: Warum wird die Varianz des Durchschnittsurteils, das heißt des Mittelwerts der Mittelwerte, im Fach nicht berücksichtigt? Wie wird die Varianz des HS-Mittelwertes bestimmt, asymptotisch oder unter Berücksichtigung des Auswahlsatzes n/N?

Wichtiger erscheint an dieser Stelle: Kann es sein, dass eine Hochschule trotz eines besseren Wertes in eine schlechtere Gruppe kommt? Um diese Frage sofort zu beantworten: ja, das ist möglich, denn Hochschulen mit gleichem Mittelwert können in verschiedenen Ranggruppen landen (vgl. HS 11, HS 15 und HS 41 in Abbildung 2), da ja auch die Streuung der Urteile berücksichtigt wird; dagegen ist erst einmal nichts einzuwenden. Eine Hochschule kann nur dann in die Schluss- bzw. in die Spitzengruppe kommen, wenn die Streuung der Urteile der Hochschule nicht zu groß ist. Das CHE sagt in seinem Arbeitspapier 119[14], in Kapitel 6.2 dazu Folgendes: „Die Ranggruppen

14 Berghoff, Sonja; Federkeil, Gero; Giebisch, Petra; Hachmeister, Cort-Denis; Hennings, Mareike; Roessler, Isabel; Ziegele, Frank: Das HochschulRanking 2009 Vorgehensweise und Indikatoren, Gütersloh, Arbeitspapier 119, Mai

haben allerdings nur eine grobe, orientierende Funktion. [...] zwischen vielen Studienbereichen der Mittelgruppe und den Extremgruppen [bestehen] keine signifikanten Mittelwertunterschiede. Die hier gewählte Ranggruppenzuordnung erlaubt aber zuverlässig die Identifikation von ‚gut' und ‚schlecht' bewertet, wenn man das Durchschnittsurteil im Fach zum Maßstab nimmt" (CHE, AP119, S. 73). Das heißt aber: ‚Gelb' ist nicht immer ‚schlechter' als ‚grün' und auch nicht immer ‚besser' als ‚rot'! Ist ‚gelb' nicht eigentlich (oder vielfach) ‚grau'? Genau diese Fragen sind nicht nur aus Bonn erhoben worden. Dazu gab es – wie gesagt – einige Erörterungen mit Mitarbeiter/inne/n des CHE, deren Ergebnisse in der LRK-Präsentation von Herrn Kollegen Ziegele vorgestellt wurden, wo man sich also des Einflusses (zu) großer Varianz auf die Gruppenzugehörigkeit angenommen hat.

4.1 Streuung der Urteile

Die Streuung der Urteile schien von Seiten des CHE zunächst vornehmlich dahingehend interpretiert zu werden, wie einig sich die Studierenden in ihren Urteilen sind. Die Streuung/Varianz hängt aber nicht nur hiervon ab, sondern (vor allem) auch von der Stichprobengröße. Damit lässt das Verfahren die Vermutung aufkommen, dass die Gruppenzuordnung unter anderem (auch) vom Rücklauf (=Stichprobengröße) abhängt und nicht (nur) von der Homogenität der Urteile, denn geringe Varianz kann Übereinstimmung signalisieren, ist aber auch bei großen Stichproben wahrscheinlicher; große Varianz ist hingegen bei kleinen Stichproben wahrscheinlicher. Zumindest also liegt (bei kleinen Stichproben u. U.) eine Konfundierung mit Heterogenität im Urteil vor! Es ist daher zu erwarten, dass Fachbereiche mit geringem Rücklauf eher der Mittelgruppe zugeordnet werden! Werden „gute" Fachbereiche durch geringen Rücklauf tendenziell bestraft, „schlechte" Fachbereiche hingegen belohnt, nämlich jeweils mit einer Zuordnung in die Mittelgruppe? Für die Stabilisierung der Gruppenzuordnung und zur Vermeidung der „Tendenz zu Gelb" ist die Festlegung einer höheren Mindeststichprobengröße dienlich. So war auch die Erhöhung auf n=15 im Mai die erste Reaktion vom CHE auf die geäußerte Kritik, wenn auch mit dieser Stichprobengröße die Argumente nicht zu entkräften sind.

4.2 Stichprobengröße und Rücklauf

Bei der Studierendenbefragung sind also die Stichprobengrößen das zentrale Problem, welches (seinerzeit) zum Beispiel Fragen bezüglich der Größe der Grundgesamtheit (N) und der Anzahl angeschriebener Studierender (n brutto) generierte: Wie groß sind der Auswahlsatz und die Rücklaufquote (n/N, bzw. n netto /n brutto) in einem bestimmten Fachbereich überhaupt? Die Anzahl der Befragungsteilnehmer (n netto) wird mittlerweile angegebenen, aber die genaue Größe der Grundgesamtheit (N), die Anzahl der angeschriebenen Studierenden n brutto (hiermit verbindet sich dann das Stichwort unit nonresponse) und auch die Anzahl der Antworten je Frage (Stichwort item nonresponse) sind weiterhin, zumindest für den Endverbraucher auf Fach- und Universitätsebene, unbekannt beziehungsweise nicht direkt einsehbar. Einige Informationen zur Anzahl der befragten Studierenden und zu Rücklaufquoten gibt es allerdings inzwischen: Es werden (maximal) 500 Studierende pro Studiengang angeschrieben (und zwar über die Universitäten). Für Bonn verhielt es sich so: Das CHE erhielt von der Universität die Studierendenzahlen der einbezogenen Studiengänge. Der Umfang der Stichprobe wurde vom CHE vorgegeben, im zuständigen Dezernat der Universität wurden die Adressen für die Stichprobe gezogen. Zur Stichprobengewinnung noch einige Informationen: Bei Diplom- und Magisterstudiengängen werden das 5. bis 12. Fachsemester berücksichtigt (in Bachelorstudiengängen das 3.-7. Fachsemester). Es werden nur Befragte in die endgültige Auswertung einbezogen, die mindestens ein Jahr an der derzeitigen Hochschule studiert haben (s. Arbeitspapier Nr. 119, S. 34). Die Adressierung und der Versand der Fragebögen an die Studierenden erfolgten von der Universität aus. Das CHE erhält (aus Datenschutzgründen) keine

2009, 90 Seiten, ISBN 978-3-939589-87-7, ISSN 1862-7188.

Adressen von Studierenden. Die Beantwortung des Fragebogens ist ausschließlich online möglich. Manche Universitäten kontaktieren sogar von sich aus n>500. Wenn ein Studiengang n<500, dann werden alle Studierenden kontaktiert. Das heißt, eine gewisse Basis für die Ermittlung einer Rücklaufquote ist gegeben und wird auch genutzt.

„Nichteinbeziehung (eines Studiengangs bei einer Rücklaufquote) unter 10% bei (einer Stichprobe von) n<30" (Statement von Herrn Kollegen Ziegele bei der LRK-NRW Tagung), wird laut CHE sogar fallweise strenger gehandhabt, denn zum Beispiel auch bei n=50 werde diese Stichproben dennoch NICHT berücksichtigt, wenn die Rücklaufquote deutlich <10% (dies mag dem Non-Responder-Problem geschuldet sein, s.u.). Das CHE kennt diese Quoten also universitätsspezifisch. In den Dokumenten werden diese Zahlen für ein bestimmtes Fach jedoch über alle Universitäten hinweg gemittelt angegeben[15]. Es ist nicht offensichtlich, wie hoch die Rücklaufquoten an den einzelnen Universitäten sind. Diese werden zumindest nicht bei der Fachbereichsdarstellung zu einem Fach dargestellt. Die Rücklaufquote variiert stark nach Fächern und Hochschultyp. In der Übersicht über den Fragebogenrücklauf der Studierendenbefragung pro Fach für Deutschland[16] ist die jeweilige durchschnittliche Rücklaufquote angegeben.

4.3 Non-Responder und Erwartungstreue

Eigentlich geht es aber nicht um Stichprobengröße und Mindestrücklaufquote, es geht um Erwartungstreue. Die Festsetzung einer Mindestrücklaufquote (also z.B. >10%) ist weder per se zielführend, noch hinreichend für die Erwartungstreue der Ergebnisse. Entscheidend ist, dass sich die Non-Responder nicht von den Respondern unterscheiden. Die Erwartungstreue der Befunde („Repräsentativität") kann aufgrund des unvermeidbaren Non-Response aber nur durch begleitende Forschung gesichert werden, welche das CHE plant. Die Gefahr ist also erkannt, aber nicht gebannt – vergleiche dazu auch die Diskussion in der Umfrageforschung, wo dies ein hochaktuelles Forschungsthema ist.

4.4 Veränderung der Ranggruppenmethode (seit 2011)[17]

Nicht verändert hat sich in 2011, dass abhängig von der Breite der Konfidenzintervalle letztendlich eine Gruppierung in drei Gruppen stattfindet: Spitzen-, Mittel- und Schlussgruppe. Neu ist seit 2011, dass nun um den bundesweiten Mittelwert zwei zusätzliche Grenzen eingezogen werden. Fachbereiche, deren Konfidenzintervall komplett zwischen diesen beiden Grenzen liegt, werden, sofern sie nicht bereits der Spitzengruppe angehören, der Mittelgruppe zugeordnet. Ist das Konfidenzintervall so breit, dass es über den Mittelwert und eine dieser Grenzen hinaus ragt, werden diese Werte nicht in das Ranking einbezogen, weil sie nicht klar einer Gruppe zuzuordnen sind. In Abbildung 3 sind HS15, HS37, HS39 und HS43 deswegen dunkelgrau; in der Regel betrifft dies insbesondere Fachbereiche mit einem Rücklauf zwischen 15 und 40, so heißt es einmal, 15 bis 30 an anderer Stelle(!?)[18].

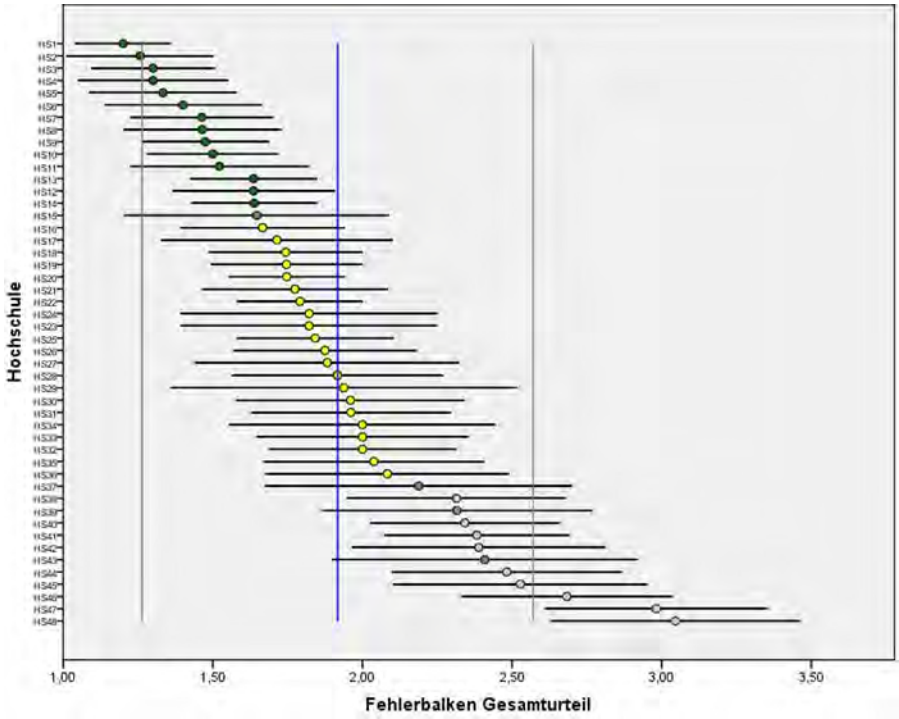
15 http://www.che-ranking.de/methodenwiki/index.php/Fragebogenr%C3%BCcklauf_der_Studierendenbefragung_pro_Fach_f%C3%BCr_Deutschland

16 http://www.che-ranking.de/methodenwiki/index.php/Fragebogenr%C3%BCcklauf_der_Studierendenbefragung_pro_Fach_f%C3%BCr_Deutschland

17 http://www.che-ranking.de/methodenwiki/index.php/Statistische_Ermittlung_von_Ranggruppen

18 http://www.che.de/cms/?getObject=318&GetName=Fehlerbalkendiagramme+f%FCr+Studierendenurteile&getLang=de

Abb. 3: Fehlerbalkendiagramm für Studierendenurteile



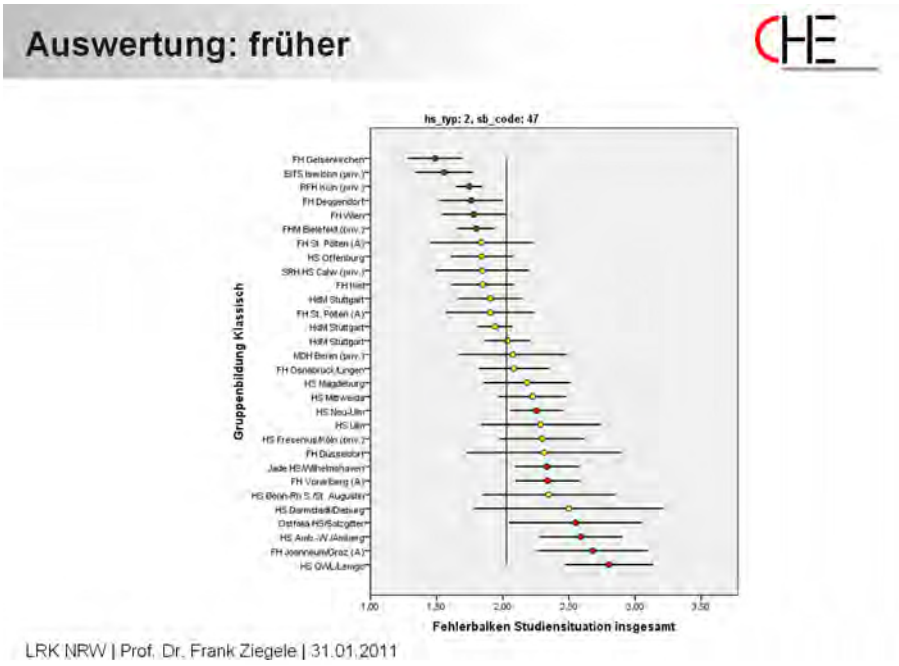
Quelle: www.che.de/methodenwiki

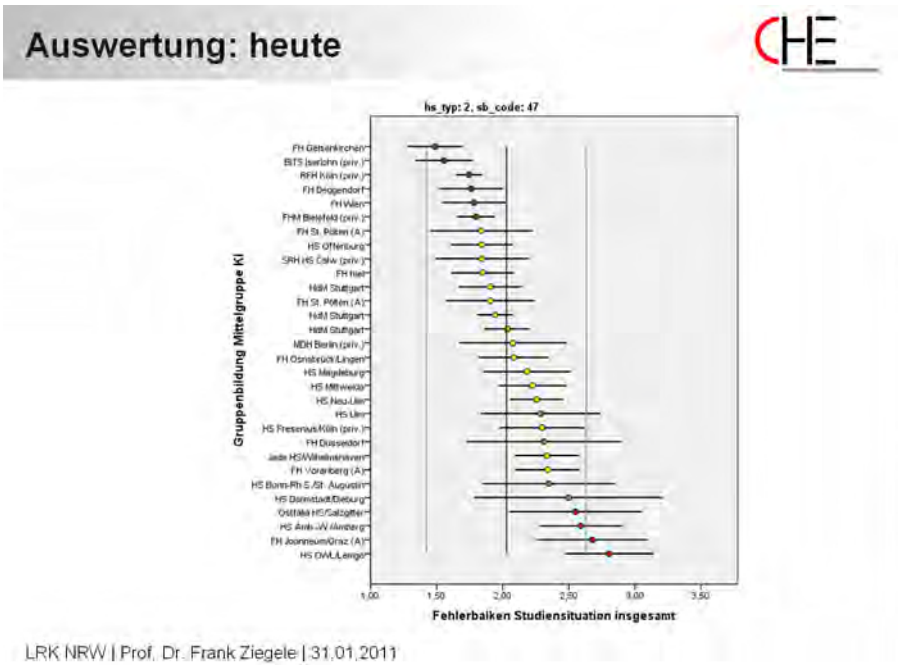
Der Bereich um den Mittelwert der Mittelwerte ist definiert als „+/-¼ Note + Durchschnitt der Standardfehler aller Einzelbeurteilungen der Fachbereiche"; diese Definition findet sich (noch) nicht im Methodenwiki, sondern wurde bei den erwähnten Gesprächen kommuniziert [19]. Mit dieser (rustikalen) Methode lassen sich Fehlrankings (besserer Mittelwert/schlechtere Gruppe etc.) entscheidend reduzieren. Die Veränderung der Ranggruppenmethode mit der Einführung einer Bandbreite um den Gesamt-Mittelwert, welche zu einer Trennung der klaren Mittelgruppe (gelb) von Fällen mit zu großen Konfidenzintervallen (die nicht gerankt werden – grau) impliziert, führt zu einer Reduzierung der Inkonsistenzen. Nicht gerankt wird auch, wenn bei Studierendenbefragung alle Werte dicht beieinander und auf hohem Niveau liegen, d.h. wenn die Spreizung der Werte kleiner 1 (z. B. zwischen 1.0 und 1.9) ist oder wenn der Schlussgruppenwert <2 ist, das heißt alle (zu) gut sind.

Im Folgenden zwei Charts aus der Ziegele-Präsentation, welche die Unterschiede zum früheren Procedere noch einmal veranschaulichen. Man achte vor allem auf die letzten 12 Fehlerbalken (von HS Neu-Ulm bis HS OWL/Lemgo): 7 davon wechseln ihre Farbe: aus rot wird dreimal gelb, aus gelb wird viermal grau!

19 Dies gilt auch für die folgenden Erläuterungen, welche in der LRK-Präsentation skizziert sind und bei den Gesprächen detailliert wurden.

Abb. 4: Fehlerbalkendiagramme - Zeitlicher Vergleich (Quelle: s. Abb. 1)





Man hat sich also durchaus des Problems angenommen, dass bei geringem Rücklauf (also großer Varianz) eine Tendenz zur Mitte – gelb – besteht, indem uneindeutig zu interpretierende Gelbs in Grau verwandelt werden, vergleiche dazu die eingangs zitierte „Bonner" Frage. Bemerkenswert zudem, dass auch einige „rote" Beurteilungen ihre Farbe zu gelb wechseln! Dennoch bleibt natürlich offen, ob und wie die Homogenität der Urteile von den Effekten der Rücklaufquoten beziehungsweise Stichprobengrößen getrennt werden kann. Die pdf-Downloads zu den Fächern enthalten jeweils eine Übersicht über die Fallzahlen zur Studierendenbefragung (ab 2010) sowie die zugehörigen Fehlerbalkendiagramme mit eingezeichnetem Mittelwert. Für 2011 werden die jeweils nach dem neuen Kriterium bei einzelnen Indikatoren nicht gerankten Fachbereiche in den Fehlerbalkendiagrammen nicht dargestellt. Die Diagramme sollen in Kürze zur Verfügung stehen.

5 Aufbereitung

Wie schon mehrfach angedeutet, wird im Zuge der Veränderung der Ranggruppenmethode von der Ampel-Symbolik Abstand genommen, um in Einzelfällen irreführende Signale durch diese Symbolik (rot = stopp) zu vermeiden und eher die positiven Empfehlungen zu betonen (Ziegele, 31. Januar 2011, LRK-NRW). Grün, aber nicht Rot, das nun durch Blau ersetzt wird, sticht jetzt (im Kontext von Gelb, Blau und Grau bzw. Weiß = nicht gerankt) wahrnehmungs- und farbpsychologisch fundiert ins Auge.

5.1 Ranking Kompakt

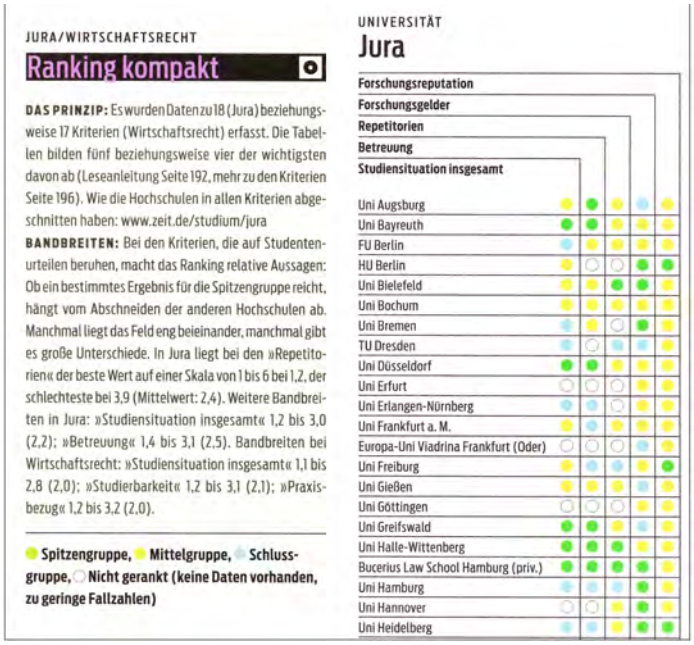
Zwei Beispiele – aus dem online-Ranking[20] und aus dem Studienführer 2011/2012 – sind in den folgenden Abbildungen gegeben, in denen die beschriebenen Modifikationen auf der „kompakten" Nutzerebene sichtbar werden. Aus grauen Fehlerbalken werden weiße Punkte/leere Kreise.

Abb. 5: ZEIT / CHE Online-Ranking 2011

Quelle: siehe Fußnote 20

20 http://ranking.zeit.de/che2011/de/rankingkompakt

Abb. 6: ZEIT Studienführer - Ranking kompakt



JURA/WIRTSCHAFTSRECHT

Ranking kompakt

DAS PRINZIP: Es wurden Daten zu 18 (Jura) beziehungsweise 17 Kriterien (Wirtschaftsrecht) erfasst. Die Tabellen bilden fünf beziehungsweise vier der wichtigsten davon ab (Leseanleitung Seite 192, mehr zu den Kriterien Seite 196). Wie die Hochschulen in allen Kriterien abgeschnitten haben: www.zeit.de/studium/jura

BANDBREITEN: Bei den Kriterien, die auf Studentenurteilen beruhen, macht das Ranking relative Aussagen: Ob ein bestimmtes Ergebnis für die Spitzengruppe reicht, hängt vom Abschneiden der anderen Hochschulen ab. Manchmal liegt das Feld eng beieinander, manchmal gibt es große Unterschiede. In Jura liegt bei den »Repetitorien« der beste Wert auf einer Skala von 1 bis 6 bei 1,2, der schlechteste bei 3,9 (Mittelwert: 2,4). Weitere Bandbreiten in Jura: »Studiensituation insgesamt« 1,2 bis 3,0 (2,2); »Betreuung« 1,4 bis 3,1 (2,5). Bandbreiten bei Wirtschaftsrecht: »Studiensituation insgesamt« 1,1 bis 2,8 (2,0); »Studierbarkeit« 1,2 bis 3,1 (2,1); »Praxisbezug« 1,2 bis 3,2 (2,0).

Spitzengruppe, Mittelgruppe, Schlussgruppe, Nicht gerankt (keine Daten vorhanden, zu geringe Fallzahlen)

UNIVERSITÄT

Jura

	Forschungsreputation	Forschungsgelder	Repetitorien	Betreuung	Studiensituation insgesamt
Uni Augsburg					
Uni Bayreuth					
FU Berlin					
HU Berlin					
Uni Bielefeld					
Uni Bochum					
Uni Bremen					
TU Dresden					
Uni Düsseldorf					
Uni Erfurt					
Uni Erlangen-Nürnberg					
Uni Frankfurt a. M.					
Europa-Uni Viadrina Frankfurt (Oder)					
Uni Freiburg					
Uni Gießen					
Uni Göttingen					
Uni Greifswald					
Uni Halle-Wittenberg					
Bucerius Law School Hamburg (priv.)					
Uni Hamburg					
Uni Hannover					
Uni Heidelberg					

Quelle: ZEIT Studienführer 2011/2012

5.2 Profil statt Ranking?

Ein weiterer Bonner Einwand war, dass man doch statt von Rankings von Profilen ausgehen sollte, die den Interessen der künftigen Studierenden bzw. Studienortwechsler am ehesten entsprächen. Universitäten sind ja geradezu gehalten, Strategien zu entwickeln, um ihre Differenziertheit, ihre Alleinstellungsmerkmale, herauszustellen und um im Wettbewerb zwischen den Hochschulen erfolgreich zu sein. Ein Ranking, das nun solche Schwerpunktsetzungen auf Seiten der Universität, aber vielleicht eben auch auf Seiten der Interessenten für ein Studium nicht berücksichtigt, passt unter Umständen auch gar nicht mehr in die moderne Hochschullandschaft. Das mag sich auch das CHE gesagt haben. Deswegen gibt es Ansätze, einen Abgleich zwischen Interessen, Erwartungen, Kompetenzen der „Klienten" und dem Studienangebot, den Rahmenbedingungen, herzustellen. Es handelt sich zum einen um das sogenannte Quick Ranking[21] (für Eilige), welches der Mehrdimensionalität der Hochschulleistungen Rechnung tragen will. In diesem online Quick Ranking können die Nutzer sich ihren Interessenaspekt wählen, z.B. „Viel Forschung", welche dann wiederum operationalisiert wird über die bekannten Indikatoren Forschungsmittel, Publikationen, Promotionen, Zitationen usw. So interessant dieser Ansatz auch sein mag, man wird wieder auf die Validität, die Triftigkeit, die Genauigkeit, die Aussagekraft dieser Indikatoren zurückgeworfen, an denen ja unter verschiedenen (technisch-methodischen) Aspekten Kritik geübt werden kann, bis hin zu der Grundsatzfrage, ob diese Indikatoren überhaupt das Profil und das Leitbild einer Universität (z. B. eben der Universität Bonn) abbilden. Dabei ist man natürlich in einem gewissen Dilemma, denn um Vergleiche zwischen Universitäten herzustellen, bedarf es eines gemeinsamen Kernes, denn eine Universität definiert sich ja nicht nur über Alleinstellungsmerkmale. Wenn als primäre Aufgabe der Universitäten erachtet wird, in Forschung und Lehre gemeinsam Wissenschaft

21 http://ranking.zeit.de/che2011/de/quickranking/show?esb=1&ab=4&hstyp=1&left_f1=450&left_f2=461 Beim „Quick Ranking" kann jeder für sein Wunschfach bis zu zehn Kriterien festlegen, die ihm selbst besonders wichtig sind. Das Ergebnis wird in einer grafischen Darstellung ausgewiesen. Die Hochschulen, die bei der Kriterienauswahl am besten abgeschnitten haben, rücken in die Mitte einer Grafik. Verändert man die Auswahl der Kriterien, können sich andere Hochschulen ins Zentrum bewegen, andere wandern an den Rand.

zu betreiben, in spannende Wissenschaft einzuführen und weiterführende wissenschaftliche Ergebnisse zu erziehen und man der Meinung ist, diese Kriterien seien an der eigenen Universität am besten erfüllt, dann hatte man sich in Bonn die Frage gestellt, ob ein solches Ranking, welches mit den gleichen Operationalisierungen an alle Universitäten herangeht, einen interessierten Studenten an den Ort führt, wo genau dies beschriebene Ziel am besten verfolgt wird. Denn, so die Fortführung des Arguments, erfolgreich im CHE-Ranking wäre dann ja insbesondere die Universität, die die Studienbedingungen ganz im Sinne der CHE-Fragen auslegte und z. B. nicht die wissenschaftliche Qualität der jeweiligen Fächer in den Vordergrund stellte (vgl. dazu die eingangs skizzierte Kritik am Inhalt des Fragenkatalogs).

Dass solche Bedenken durchaus ernst genommen werden, mag man daran ablesen, dass die Profilbildung der verschiedenen Fächer durch fachspezifische Beiräte gewährleistet werden soll. Darauf ist in seiner Präsentation vom 31.01.2011 Herr Kollege Ziegele auch an mehreren Stellen eingegangen, zum Beispiel bezüglich der Kriterien für Internationalisierung und hinsichtlich von Facetten für die Profilbildung (z.B. in der Medienwissenschaft). Fachbereichsbefragungen sollen – nach Inhaltsanalyse der Texte – die profilbildenden Merkmale gemeinsam mit den Hochschulen, erarbeitet für die nächste Runde, liefern. Dieser Profilbildungansatz soll dann wohl auch bei „Mein Ranking“[22] – einer weiteren interaktiven Ranking-Form – Eingang finden.

6 Politische Dimension des Rankings

Ohne Frage, die Hochschulen brauchen eine Präsentation der eigenen Universität in Forschung und Lehre, die interessierten Studierenden, aber auch internationalen Gastwissenschaftlern und potentiell neuen Kolleginnen und Kollegen eine differenzierte Entscheidungsgrundlage bietet.

Welche Rolle spielt dabei das CHE-Ranking auch und gerade im Sinne der Aufwand-Nutzen-Relation für die Hochschulen?
Warum dieser Aufwand (z. B. Versenden der Adressen, Infos versenden und prüfen…) für Dritte, für deren Publikationserfolg: >100 000 verkaufte Exemplare à 7.95€ pro Heft und Aber-Millionen Seitenaufrufe?
Auf den andere Seite muss man sich natürlich klar machen, dass ein so auflagenstarker Studienführer eine hohe wissenschafts-politische Brisanz hat und nicht nur dem Entertainment dient. Verbessert sich durch die Veränderungen von Seiten des CHE, die hier beschrieben worden sind, die Aufwand-Nutzen-Relation für die Hochschulen in dem Sinne, dass aufgrund dessen auch die passenden Studenten an die entsprechende, z.B. die Bonner Universität, kommen?

Lässt sich aber eine solche Passung nicht auch durch professionelle Fächerwahl-Tests in Form von Online-Self-Assessment (OSA) als eigene „Werbung“ für den Standort Bonn erreichen? So macht die Bonner Universität durch ihr Studieninformationssystem als zweistufigem Eingangsportal eigene Werbung für den Standort.
Die erste Ebene gibt allgemeine Informationen zum Studium in Bonn und die zweite Ebene gibt spezifische Informationen zum Studiengang, zum Studienfach und bietet Zugang zu den fachspezifischen OSAs und versucht damit, dem veränderten Orientierungsbedarf bei Studieninteressierten durch Ausdifferenzierung des Studienangebotes nach der Bologna-

22 http://www.che.de/methodenwiki/index.php/Mein_Ranking
In der Internetversion des HochschulRankings besteht die Möglichkeit, sich entsprechend eigener Prioritäten interaktiv ein persönliches Ranking zu erstellen, in dem individuell bis zu 5 verschiedene Indikatoren ausgewählt und in ihrer Bedeutung gewichtet werden können. In der Funktion „Mein Ranking“ wird der Anwender Schritt für Schritt durch die verschiedenen Auswahlmöglichkeiten geführt.

Reform Rechnung zu tragen. Es geht also vor allem um die Vermittlung von Erwartungen, um die Verdeutlichung des Anspruchsniveaus, um die Gewinnung qualifizierter Studierender durch Selbstselektion, um letztendlich damit die Erfolgsquote im Studium zu erhöhen und die Abbrecher- oder Fachwechslerquote zu reduzieren.

Solche online-Plattformen bieten auch die Möglichkeit, überregionale und internationale Rekrutierung von Studierenden zu betreiben und zielgruppenadäquate Ansprache für den Erstkontakt mit einer Universität auszuüben. [23]

Bedenkenswert allerdings ist, dass im CHE-ZEIT-Studienführer (natürlich) auch die Links zu den entsprechenden Self-Assessment-Portalen (z.B. Bochum, des Landes Baden-Württemberg, Freiburg, RWTH Aachen, Nordverbund) aufgeführt sind, das Bonner OSA (http://www3.uni-bonn.de/studium/studienangebot/studienscout-academicus) jedoch (noch) nicht aufgeführt wird.

7 Fazit und Nachtrag

Dem Senat der Universität Bonn wurden die Reaktionen des CHE auf die kritischen (Bonner) Argumente durch den Erstautor dieses Beitrags am 9. Juni 2011 vorgestellt. Ob die bibliometrischen Analysen lege artis sind und eine hinreichende Differenzierung zwischen den Fachkulturen gewährleisten, wurde genauso diskutiert wie die zum Wissenschaftsbezug hinzugenommenen Fragen bei der Studierendenbefragung im Sinne einer hinreichenden Stärkung dies Aspekts. Bei der Ranggruppenzuordnung wurde festgestellt, dass das CHE sich bewegt hat, was die Rücklaufquote, die Stichprobengröße, die Ampelsymbolik (hinter deren Änderung ja auch ein veränderter statistischer Ansatz steht) angeht. Dass die Fachbeiratsmitglieder, welche ja doch gewisse „politische" Funktionen haben, nicht bekannt gemacht werden, ist weiterhin nicht auf Akzeptanz gestoßen. Der Senat hat dem Rektorat mit 11:6 Stimmen bei 4 Enthaltungen empfohlen, beim CHE-Ranking wieder einzusteigen.

Das Rektorat ist auf seiner Sitzung am 28. Juni 2011 dieser Empfehlung gefolgt:

> *„Die Universität Bonn wird künftig wieder an den Rankings des Centrums für Hochschulentwicklung (CHE) teilnehmen. Das hat das Rektorat der Universität jetzt beschlossen. Der Entscheidung war eine mehrjährige Diskussion vorausgegangen. Weil das CHE substantielle Korrekturen an Darstellung und Methodik ihres Rankings vorgenommen habe, stehe der erneuten Teilnahme der Universität Bonn am CHE-Ranking nun nichts mehr im Wege, begründete das Rektorat seine Entscheidung"*[24].

23 Vielleicht ist dies gar der erste Schritt der Bindung der künftigen Studierenden an die Universität, die sie auf der Basis und aufgrund eines solchen Angebotes gewählt haben. s. a. Rudinger, G & Hörsch, K. (Hg.) (2009): Self-Assessment an Hochschulen: Von der Studienfachwahl zur Profilbildung. Applied Research in Psychology and Evaluation, Band 4. Bonn University Press. Göttingen: V&R unipress.

24 Aufmacher der Pressemitteilung pm177-11 zum Wiedereintritt der Universität Bonn ins CHE-Ranking: http://www3.uni-bonn.de/Pressemitteilungen/177-2011

Verzeichnis der Autorinnen und Autoren

Prof. Dr. Eva Barlösius
Leibniz Universität Hannover
Institut für Soziologie und Sozialpsychologie
Schneiderberg 50; 30167 Hannover
www.ish.uni-hannover.de
http://eva.barloesius.phil.uni-hannover.de

Dr. Silke Gülker
WZB Wissenschaftszentrum Berlin für Sozialforschung
Forschungsgruppe Wissenschaftspolitik
Reichpietschufer 50; 10785 Berlin
www.wzb.eu
http://www.wzb.eu/de/personen/silke-guelker

Dr. Norbert Hilger
Zentrum für Evaluation und Methoden (ZEM) der Universität Bonn
Oxfordstraße 15; 53111 Bonn
www.zem.uni-bonn.de
www.zem.uni-bonn.de/ueber-uns/mitarbeiter/norbert-hilger

Prof. Dr. Stefan Hornbostel
iFQ – Institut für Forschungsinformation und Qualitätssicherung
Schützenstraße 6a; 10117 Berlin
www.forschungsinfo.de
http://www.forschungsinfo.de/Mitarbeiter/mit_horn.asp

Dr. Thamar Klein
iFQ – Institut für Forschungsinformation und Qualitätssicherung
Schützenstraße 6a; 10117 Berlin
www.forschungsinfo.de
http://www.forschungsinfo.de/Mitarbeiter/mit_klein.asp

Dr. Wilhelm Krull
Volkswagenstiftung
Kastanienallee 35; 30519 Hannover
www.volkswagenstiftung.de
http://www.volkswagenstiftung.de/stiftung/generalsekretaer.html

Prof. Michèle Lamont
Harvard University, Department of Sociology
510 William James Hall
33 Kirkland Street; Cambridge, MA 02138
www.wjh.harvard.edu/soc
http://www.wjh.harvard.edu/soc/faculty/lamont/

Prof. Dr. Axel Michaels
Universität Heidelberg, Südasien-Institut
Im Neuenheimer Feld 330; 69120 Heidelberg
www.sai.uni-heidelberg.de
http://www.sai.uni-heidelberg.de/abt/IND/mitarbeiter/michaels/michaels.php

Meike Olbrecht
iFQ – Institut für Forschungsinformation und Qualitätssicherung
Schützenstraße 6a; 10117 Berlin
www.forschungsinfo.de
http://www.forschungsinfo.de/Mitarbeiter/mit_olbrecht.asp

Prof. Dr. Georg Rudinger
Zentrum für Evaluation und Methoden (ZEM) der Universität Bonn
Oxfordstraße 15; 53111 Bonn
www.zem.uni-bonn.de
http://www.zem.uni-bonn.de/ueber-uns/mitarbeiter/prof.-dr.-georg-rudinger

Dr. Dagmar Simon
WZB Wissenschaftszentrum Berlin für Sozialforschung
Forschungsgruppe Wissenschaftspolitik
Reichpietschufer 50; 10785 Berlin
www.wzb.eu
http://www.wzb.eu/de/personen/dagmar-simon

Dr. Marc Torka
WZB Wissenschaftszentrum Berlin für Sozialforschung
Forschungsgruppe Wissenschaftspolitik
Reichpietschufer 50; 10785 Berlin
www.wzb.eu
http://www.wzb.eu/de/personen/marc-tor