Röbbecke, Martina (Ed.); Simon, Dagmar (Ed.)

Working Paper

Qualitätsförderung durch Evaluation? Ziele, Aufgaben und Verfahren von Forschungsbewertungen im Wandel

WZB Discussion Paper, No. P 99-003

Provided in Cooperation with:
WZB Berlin Social Science Center

Suggested Citation: Röbbecke, Martina (Ed.); Simon, Dagmar (Ed.) (1999) : Qualitätsförderung durch Evaluation? Ziele, Aufgaben und Verfahren von Forschungsbewertungen im Wandel, WZB Discussion Paper, No. P 99-003, Wissenschaftszentrum Berlin für Sozialforschung (WZB), Berlin

This Version is available at:
https://hdl.handle.net/10419/50253

P 99 – 003

QUALITÄTSFÖRDERUNG DURCH EVALUATION?

Ziele, Aufgaben und Verfahren
von Forschungsbewertungen im Wandel

MARTINA RÖBBECKE ♦ DAGMAR SIMON (HG.)

DOKUMENTATION DES WORKSHOPS

VOM 20. UND 21. MAI 1999

Wissenschaftszentrum Berlin für Sozialforschung gGmbH (WZB)
Reichpietschufer 50, D-10785 Berlin

Dr. Martina Röbbecke
Tel. 030 - 254 91 -534
e-mail: roebbecke@medea.wz-berlin.de

Dr. Dagmar Simon
Tel. 030 - 254 91 -588
e-mail: dsimon@medea.wz-berlin.de

Wissenschaftszentrum Berlin für Sozialforschung
Reichpietschufer 50
D-10785 Berlin

Zusammenfassung

Dieses Paper dokumentiert einen Workshop, der sich mit Evaluationen außeruniversitärer, staatlich finanzierter Forschungseinrichtungen auseinandergesetzt hat. Im Unterschied zu den Hochschuldebatten über Evaluationen, die bereits in den achtziger Jahren eingesetzt haben und die sich nicht nur in einer Fülle von Publikationen, sondern auch in einer ansehnlichen Zahl von Reformprojekten niederschlagen, finden im außeruniversitären Sektor Diskussionen in einem vergleichbaren Maß nicht statt. Dabei können viele Forschungseinrichtungen auf langjährige Evaluationserfahrungen zurückblicken: sei es auf interne Bewertungsverfahren durch Institutsbeiräte, sei es auf externe Evaluierungen durch den Wissenschaftsrat. Mit dem Workshop war intendiert, eine Debatte über Ziele, Aufgaben, Verfahren und Instrumente von Forschungsevaluationen aus den unterschiedlichen Perspektiven von Wissenschaftspolitikern, Evaluationsspezialisten sowie Praktikern aus Hochschulen und außeruniversitären Forschungseinrichtungen zu führen.

Konkreten Anlaß für diesen Workshop gab das am WZB durchgeführte Forschungsprojekt "Institutionelle Selbstbeobachtung als Steuerungsinstrument für außeruniversitäre Forschungseinrichtungen?". Gegenstand des Projekts sind Forschungseinrichtungen der Wissenschaftsgemeinschaft G. W. Leibniz (WGL). Besondere Aufmerksamkeit erfahren dabei die institutsspezifischen Strukturen und Prozesse, denn eine Qualitätsbewertung und Qualitätsförderung der Forschung kann nur dann überzeugend gelingen, wenn die Instrumente und Verfahren der Evaluation die jeweiligen Aufgaben und Arbeitsweisen und damit das spezifische Profil einer Einrichtung erfassen. Jede Konzeption eines Evaluationsverfahrens steht daher vor der Herausforderung, den hinsichtlich ihrer Forschungsorientierungen, der disziplinären Ausrichtungen, der Institutsstrukturen sowie der Außenanforderungen sehr heterogenen Einrichtungen gerecht werden zu müssen. Auf die Frage, ob und wie es gelingen kann, Evaluationen in diesem Spannungsfeld als Verfahren der Qualitätsförderung zu etablieren, versuchen die Beiträge des Workshops erste Antworten zu geben.

Inhaltsverzeichnis

Einleitung

Qualitätsförderung durch Evaluation? S. 7

Ulrike Felt

Evaluation im wissenschaftspolitischen Kontext S. 11

Ekkehard Nuissl von Rein

Unterschiedliche Aufgaben – gemeinsame Ziele?
Entwicklung und Bewertung der Leibniz-Institute S. 31

Martina Röbbecke

Einheitlichkeit oder Eigensinn?
Angemessene Indikatoren für heterogene Forschungseinrichtungen S. 46

Stefan Hornbostel

Welche Indikatoren zu welchem Zweck: Input, Throughput, Output S. 55

Dagmar Simon

Wer evaluiert zu welchem Zweck was?
Anmerkungen zu Zielen und Verfahren der Selbstevaluation
in außeruniversitären Forschungseinrichtungen S. 73

Jürgen Lüthje

Impulse und mögliche Parameter für die Forschungsevaluation S. 81

Adrian C. L. Verkleij

Self-evaluation and External Review S. 87

Ulrich Teichler

Hochschulevaluation und Hochschulmanagement
im internationalen Vergleich – einige Thesen S. 100

Programm des Workshops S. 114

Teilnehmerinnen und Teilnehmer S. 115

Einleitung

Qualitätsförderung durch Evaluation?

Eine bloß rhetorische Frage? Sind die Ziele von Evaluationen im Wissenschafts- und Forschungsbereich nicht (eindeutig) definiert? Und wer würde ernsthaft gegen Qualitätsförderung als Zieldefinition von Evaluation sprechen? Sicherlich niemand, dennoch vermitteln die zentralen Begriffe in der gegenwärtigen Debatte über die mit Evaluationen verbundenen Zielvorstellungen – wie Qualitätskontrolle, Qualitätssicherung und Qualitätsförderung, Rechenschaftslegung, Transparenz und Flexibilitätssicherung, um nur einige zu nennen – nicht gerade Klarheiten.

Welche Unterschiede bestehen etwa zwischen Qualitätssicherung und Qualitätsförderung, und lassen sich diese Ziele mit anderen – wie beispielsweise der "accountability" – verbinden? Ist es überhaupt möglich, mit Evaluationen eine Qualitätsförderung zu erreichen, oder sind diese Erwartungen an ein Verfahren, das vorrangig der Bewertung erbrachter Leistungen dient, nicht viel zu hoch? Führen der Erfolgsdruck und möglicherweise empfindliche finanzielle Konsequenzen nicht dazu, daß die zu bewertenden Institutionen ihre Stärken herausstreichen und die Schwächen verdecken? Mit welchen Indikatoren läßt sich die Qualität der erbrachten Leistungen angemessen bewerten, und wie sollte ein Evaluationsverfahren gestaltet sein, das eine Qualitätsförderung ermöglicht und unterstützt?

Wir haben versucht, diese und andere Fragen im Rahmen eines Workshops zu beantworten, der am 20. und 21. Mai 1999 am Wissenschaftszentrum Berlin für Sozialforschung (WZB) stattfand. Im Zentrum des Workshops, den wir hier dokumentieren, standen Evaluationen außeruniversitärer, staatlich finanzierter Forschungsinstitute. Im Unterschied zu den Hochschuldebatten über Evaluationen, die bereits in den achtziger Jahren eingesetzt haben und die sich nicht nur in einer Fülle von Publikationen, sondern auch in einer ansehnlichen Zahl von Reformprojekten niederschlagen, finden im außeruniversitären Sektor Diskussionen in einem vergleichbaren Ausmaß nicht statt. Dabei können viele Forschungsinstitute auf langjährige Evaluationserfahrungen zurückblicken: sei es auf interne Bewertungsverfahren durch wissenschaftliche Institutsbeiräte, sei es auf externe Evaluierungen durch den Wissenschaftsrat.

Auch eine Diskussion und ein Erfahrungsaustausch zwischen den Hochschulen und den außeruniversitären Forschungseinrichtungen ist noch nicht so recht zustande gekommen. Zwar gibt es nicht zu vernachlässigende Unterschiede der verschiedenen Einrichtungen, beispielsweise zwischen dem Gegenstand der Evaluationen – Studium und Lehre auf der einen, Forschung auf der anderen Seite – sowie zwischen den institutionellen und politischen Rahmenbedingungen. Es gibt aber auch Gemeinsamkeiten in Hinblick auf zu bewältigende gesellschaftliche Forderungen nach Transparenz und Rechenschaftslegung, bei der Suche nach angemessenen Indika-

toren, bei Verfahrensfragen und insbesondere bei der Realisierung von – zumindest teilweise – gemeinsamen Zielsetzungen.

Damit ist auch eine mit diesem Workshop verbundene Absicht markiert: wir wollten eine Debatte über Ziele, Aufgaben und Instrumente von Forschungsevaluationen aus der unterschiedlichen Perspektive verschiedener Akteure führen. Daher haben wir Wissenschaftspolitiker, Evaluationsspezialisten sowie Praktiker aus Hochschulen und außeruniversitären Forschungseinrichtungen eingeladen.

Konkreten Anlaß für die Einrichtung dieses Diskussionsforums gab das am WZB durchgeführte Forschungsprojekt "Institutionelle Selbstbeobachtung als Steuerungsinstrument für außeruniversitäre Forschungseinrichtungen?". Dieses vom Stifterverband für die Deutsche Wissenschaft geförderte Projekt wird von Friedhelm Neidhardt, Martina Röbbecke und Dagmar Simon bearbeitet; im Rahmen des Workshops wurden erste Ergebnisse präsentiert und zur Diskussion gestellt. In einer Phase, in der unter dem Schlagwort "New Public Management" über neue Managementpraktiken und Qualitätspolitiken in den öffentlichen Verwaltungen nachgedacht wird, konzentriert sich das Projekt auf Instrumente und Verfahren interner Bewertungen. Vorrangig interessieren uns Formen der Selbstbeobachtung und Selbstkontrolle, die eine Reflektion der Leistungsfähigkeit im Sinne einer Stärken-Schwächen-Analyse erlauben und die Selbststeuerungsfähigkeit einer Forschungseinrichtung fördern.

Forschungsgegenstand des Projekts – und Ort zahlreicher Gespräche mit verschiedenen Mitarbeiterinnen und Mitarbeitern – sind fünf Institute der "Blauen Liste" beziehungsweise der Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz (WGL), zu der sich die Mehrzahl der Blaue Liste-Einrichtungen zusammengeschlossen hat. Besondere Aufmerksamkeit erfahren dabei die institutsspezifischen Strukturen und Prozesse, denn eine Qualitätsbewertung und Qualitätsförderung der Forschung kann nur dann überzeugend gelingen, wenn die Instrumente und Verfahren der Evaluation die jeweiligen Aufgaben und Arbeitsweisen und damit das spezifische Profil einer Einrichtung erfassen. Jede Konzeption eines Evaluationsverfahrens steht daher vor der Herausforderung, den hinsichtlich ihrer Forschungsorientierungen, der disziplinären Ausrichtungen und der Institutsstrukturen sehr heterogenen Einrichtungen gerecht werden zu müssen. Ulrike Felt hat die vielfältigen Anforderungen an Evaluationen, die sowohl durch einrichtungsspezifische Besonderheiten wie durch externe Erwartungen gekennzeichnet sind, wie folgt zusammengefaßt: "Evaluation muß somit im Spannungsfeld zwischen Qualität der Einrichtung, von außen entgegengebrachtem Vertrauen, veränderten Erwartungshaltungen, limitierten und zu legitimierenden Budgets, Verschiebungen in und zwischen Forschungsfeldern und vielem mehr verortet werden" (s. S. 11). Ob und wie es gelingen kann, Evaluationen in diesem Spannungsfeld als Verfahren der Qualitätsförderung zu etablieren, war eine zentrale Frage unseres Workshops.

Zu den einzelnen Beiträgen:

Im ersten Beitrag diskutierte *Ulrike Felt* von der Universität Wien zunächst, welche Erwartungen sich unter veränderten wissenschaftspolitischen Rahmenbedingungen an Evaluationen richten, und befaßte sich mit den Schwierigkeiten, die Qualität wissenschaftlicher Leistungen "festzumachen" – Qualität erscheint aus ihrer Perspektive als ein "moving target". Von der Referentin wurden verschiedene Evaluationstypen und die damit verbundenen Zielvorstellungen vorgestellt. In diesem Zusammenhang arbeitete sie die Bedeutung des lokalen Kontexts heraus, der im entscheidendem Maß die Praxis der Evaluation definiert.

Die Evaluation durch den Wissenschaftsrat stand im Mittelpunkt des Beitrages von *Ekkehard Nuissl von Rein*, dem Direktor des Deutschen Instituts für Erwachsenenbildung und wissenschaftlichen Vizepräsidenten der WGL. Seine Ausführungen machten insbesondere die Problematik eines weitgehend einheitlichen Kriteriensatzes für heterogene Institutstypen deutlich. Er warf das Grundsatzproblem auf, daß die Kriterien einer angemessenen Bewertung erbrachter Leistungen aufgabenabhängig deduziert werden müssen, dies jedoch mit einem fachübergreifenden Beurteilungssystem in Konflikt gerät.

Unter dem Themenkomplex "Input, Throughput, Output: Welche Indikatoren für welchen Zweck?" setzte sich zunächst *Martina Röbbecke* vom WZB mit der Frage auseinander, welche Bewertungskriterien der Vielfalt des Aufgaben- und Leistungsspektrums von WGL-Instituten angemessen sind. Sie unterschied zwischen Kennzahlen, Prozeß- und Strukturmerkmalen sowie Wissenschaftsindikatoren und machte insbesondere deutlich, daß für alle Leistungsbereiche einer Einrichtung jeweils adäquate Instrumente der Bewertung entwickelt werden müssen, wenn eine Konzentration auf einen bestimmten Forschungstypus und eine Akademisierung der Institute vermieden werden soll. *Stefan Hornbostel* vom Centrum für Hochschulentwicklung stellte in seinem Beitrag die generelle Ziel- und Aufgabengebundenheit von Indikatoren heraus, erörterte die Verwendungsmöglichkeiten von Indikatoren und schilderte verschiedene "Fallstricke und Fettnäpfchen". Beispielsweise kontrastierte er einheitliche Indikatorensets mit den unterschiedlichen (disziplinären) Wissenschaftspraktiken und Konventionen: so gibt es zum Beispiel erhebliche Unterschiede in der Publikationspraxis zwischen den Fächern Jura und Physik.

Der anschließende Veranstaltungsteil stand unter dem Thema "Verfahren der Selbstevaluation von Forschung". Einleitend arbeitete *Dagmar Simon* vom WZB die komplexen Aufgaben- und Forschungsfelder der Blaue Liste-Institute heraus, die nicht nur durch die kognitiven Strukturen bestimmt sind, sondern auch durch unterschiedliche externe Anforderungen und Kooperationserfordernisse. Deren Integration erfordert besondere strukturelle und forschungsorganisatorische Lösungen – ein Feld, auf dem erheblicher Evaluationsbedarf besteht. Anschließend entwickelte *Jürgen Lüthje*, Präsident der Universität Hamburg, Evaluationen als eine Methode qualitativer Universitätsentwicklung. Er präsentierte das Evaluationskonzept des Verbundes Nord-

deutscher Universitäten, das auf eine Verbindung von Selbstevaluationen und externen Evaluationen setzt. Dieser Ansatz von Qualitätsentwicklung erlaubt seiner Meinung nach den Fachbereichen eine Stärken-Schwächen-Analyse sowie den Fächern eine Selbstkorrektur.

Seine Darstellung wurde ergänzt um einen Bericht über die Erfahrungen in den Niederlanden, deren Evaluationsansatz erheblichen Einfluß auf die bundesdeutschen Diskussionen genommen hat. *Adrian Verkleij* vom Center for Higher Education Policy Studies ging insbesondere auf den Zusammenhang von Selbstevaluationen und externen Evaluationen ein. Die Organisation und Durchführung der externen Evaluation habe erheblichen Einfluß auf das Verfahren der Selbstevaluation; vor allem könne über externe Evaluationen – die in den niederländischen Universitäten einer Selbstevaluation folgen – gesichert werden, daß die Ergebnisse der Selbstevaluation zu konkreten Verbesserungsvorschlägen und einem institutionellen Wandel führen.

In seinem Beitrag über Hochschulevaluation und Hochschulmanagement erläuterte *Ulrich Teichler* vom Wissenschaftlichen Zentrum für Berufs- und Hochschulforschung, daß im internationalen Vergleich keine Entwicklung zu einem mehr oder weniger einheitlichen Hochschulevaluationssystem identifizierbar ist, sondern eine kaum überschaubare Vielfalt vorherrscht. Die Ausbreitung von Hochschulevaluationen gehe zudem einher mit einer wachsenden Komplexität der Machtverhältnissen, die – so seine skeptische Einschätzung – das komplexe Aufgabenfeld des Hochschulmanagements zur "mission impossible" werden läßt.

Bedauerlicherweise mußte Friedhelm Neidhardt seine Teilnahme an dem Workshop wegen Erkrankung absagen. Ein besonderer Dank geht daher an Meinolf Dierkes vom WZB, der sich spontan bereit erklärt hat, die Veranstaltung einzuleiten, und dabei auf (schon fast) historische Erfahrungen der "Selbstbeobachtung" rekurrieren konnte. Ebenfalls bedanken wir uns bei Helmut Wollmann von der Humboldt-Universität, der sich kurzfristig bereit erklärte, die Veranstaltung zu resümieren. Das Engagement der Vortragenden und die offenen, kritischen Diskussionen der Teilnehmerinnen und Teilnehmern – sowie nicht zuletzt ein warmer Sommerabend – waren wichtige Voraussetzungen eines anregenden Workshops.

Ulrike Felt

Evaluation im wissenschaftspolitischen Kontext[1]

Evaluation und ihre wissenschaftspolitische Bedeutung sind in den letzten Jahrzehnten immer häufiger aus ganz unterschiedlichen Positionen heraus diskutiert worden, und ich möchte daher den folgenden Ausführungen eine Skizze meines eigenen Grundverständnisses darüber voranstellen. Dies ist wesentlich, um meine Auswahl und Lektüre bestehender Literatur, aber auch bestimmte Schwerpunktsetzungen besser verstehen zu können. Ich verstehe Evaluation als einen Gesamtprozeß, in dem es um Methoden und deren Anwendung geht, um Interpretationsmöglichkeiten von Datenmaterial, um Entscheidungsstrukturen, institutionelle Rahmenbedingungen, Ablaufmuster, Akteurskonstellationen, Erwartungshaltungen, um Kopplungsmechanismen zwischen verschiedenen Teilsystemen des Wissenschaftssystems, aber auch zu gesellschaftlichen Teilbereichen. Es ist zwar wesentlich und legitim, diese unterschiedlichen Perspektiven und Bereiche aus Gründen der Überschaubarkeit getrennt zu diskutieren und zu analysieren, man sollte sich aber immer bewußt sein, daß die genannten Aspekte immer für die jeweils anderen maßgeblich mitgestaltend und ein Grenzen bildender Kontext sind. Evaluation ist also ein komplexer, vielschichtiger Aushandlungsprozeß, und als solchen möchte ich ihn auch jetzt aus unterschiedlichen Blickwinkeln beleuchten.

Beinahe vier Jahrzehnte sind verstrichen, seit die Evaluationsdiskussion in den USA der frühen 60er Jahre ihren Ausgang genommen hat. Evaluation hat sich in dieser Zeit sowohl in den USA, aber auch in Europa gerade durch ihre Koppelung mit wissenschaftspolitischen Entscheidungen beinahe zu einem eigenen „business sector“ (House 1993) entwickelt. Auch in der aktuellen internationalen Diskussion über die Reform der Forschungs- und Universitätssysteme ist die Auseinandersetzung um einen Qualitätsbegriff ins Zentrum gerückt und zu einem Kristallisationspunkt der Interessen und Erwartungen der unterschiedlichen Akteursgruppen geworden.[2] Damit rückt ein ganzer Komplex von Fragen in den Vordergrund: Wie kam es dazu, daß Universitäten, Forschungsinstitute oder Wissenschaftsprogramme

[1] Neben der Beschäftigung mit Evaluation im Rahmen meiner Forschungsarbeit hat mein mehrjähriges Mitwirken bei der Gestaltung von konkreten Evaluationsmaßnahmen für die Universität Wien meinen Blick auf diesen Themenkomplex stark geprägt. Die Universität Wien befindet sich derzeit in der Phase grundlegender, weichenstellender Entscheidungen in diesem Bereich, die in enger Verbindung mit ersten „Experimenten“ sowohl im Bereich Lehre als auch im Bereich Forschung gesehen werden müssen. – Für Überlegungen in diesem Zusammenhang siehe etwa auch Felt (1999). Insbesondere die Abschnitte 5 und 6 sind in leicht veränderter Form übernommen worden.

[2] Dies ist – um ein Beispiel herauszugreifen – derzeit sehr klar in den Diskussionen um die Reform des österreichischen Universitätssystems zu erkennen. Eine der zentralen Neuerungen im neuen Universitätsorganisationsgesetz (UOG 93) ist neben einer Autonomisierung auch die Einführung von regelmäßigen Evaluationen in Forschung und Lehre.

immer öfter Bewertungen unterzogen werden oder sich bemüßigt fühlen, selbst solche Leistungsprüfungen durchzuführen? Warum scheinen die bis dahin bestehenden Wissenschaftsstrukturen, die ja bereits eine Fülle von qualitätssichernden Mechanismen eingebaut haben, der heutigen Situation nicht mehr gerecht zu werden? Und schließlich: Wer sind die an diesen Hinterfragungsprozessen beteiligten Akteure, worin bestehen ihre Motive für eine Beteiligung und welche Erwartungen richten sie an den vermehrten Einsatz von Evaluationen?

Ein Vergleich bzw. das Aufzeigen der unterschiedlichen im europäischen Raum koexistierenden Evaluationskulturen – was ich in diesem Beitrag nur in Ansätzen unternehme – könnten einen wesentlichen Schritt zu einer Beantwortung dieser Fragen beitragen.[3]

Das Verhältnis zwischen wissenschaftlichen Einrichtungen und Gesellschaft konnte lange Zeit auf einem Vertrauensprinzip aufbauen: Vertrauen in wissenschaftliche Kompetenz und Expertise, in eine systemintern verankerte Motivation, die Qualität der eigenen Produkte aufrechtzuerhalten, und zum Teil natürlich in Institutionen oder Disziplinen, die über Jahrzehnte hinweg ein Netz von Kontrolleinrichtungen mehr oder weniger formaler Natur errichtet hatten. Seit dem Zweiten Weltkrieg haben sich sowohl die Produktionsbedingungen von wissenschaftlichem Wissen[4] als auch die gesellschaftlichen Kontexte, in die dieses Wissen eingebettet ist/wird, radikal verändert. Gründe für diese wachsende Fremd- und Selbstkontrolle liegen einerseits in einem nahezu exponentiellen Anwachsen – in Anzahl, Umfang (finanziell und personell) und Zeithorizont – der Forschungsaktivitäten, sowie in einer voranschreitenden Technologisierung der Forschung und in den damit verbundenen Kosten (Price 1963 und 1974; Galison und Hevly 1991; Cozzens u. a. 1990). Andererseits sind höhere Anforderungen an die Wissenschaft als ein Dienstleistungssystem für die Gesellschaft auszumachen, wobei gleichzeitig das kritische Bewußtsein, oder besser gesagt, eine ambivalente Haltung gegenüber Wissenschaft und Technik und deren gesellschaftlichen Auswirkungen im Zunehmen begriffen ist.

Weiter akzentuiert wird diese Veränderung durch einen teilweisen Rückzug des Staates aus dem Bereich der *Higher Education*, ein Rückzug, welcher vielfach als Autonomisierung, d. h. verstärkte Selbstbestimmung des universitären Sektors, rhetorisch positioniert wird. Werte und Praktiken, die lange Zeit unhinterfragt waren, werden in diesem Prozeß gewissermaßen zur Disposition gestellt und von den

[3] Es gibt bislang sehr wenige Arbeiten, die sich im Detail mit der Frage der verschiedenen nationalen Evaluationskulturen und ihren Auswirkungen auf die respektiven Wissenschaftssysteme im Detail auseinandersetzen. Auch die Kritik an Evaluation setzt meist auf einer relativ abstrakten Ebene an und geht nicht darauf ein, daß die Probleme meist im Zusammenwirken verschiedener Systemelemente zu suchen sind. Ein solcher Zugang würde ein besseres Verständnis der Unterschiede zwischen den nationalen Wissenschaftssystemen mit sich bringen.

[4] Siehe etwa die Ausführungen von Gibbons u. a. (1994) über die veränderte Form der Wissensproduktion.

gesellschaftlichen Akteuren zum Teil als „unzeitgemäß“ eingestuft und eine Abkehr von ihnen eingefordert. Die direkten Folgen sind also sowohl erhöhter Rechtfertigungsdruck von Seiten der Öffentlichkeit als auch zunehmender Wettbewerbsdruck innerhalb des Wissenschaftssystems. Es schien daher notwendig und angebracht, sowohl Kriterien und Prozesse der Beurteilung, ob und welche Art von Forschung förderungswürdig sei, zu erarbeiten als auch Mechanismen einer „begleitenden Kontrolle“ zu etablieren. Verstärkte Evaluationsmaßnahmen könnte man in einer solchen Interpretation als „Preis“ für eine neue Freiheit sehen: Die Erweiterung der Freiräume ist eng gekoppelt mit der Gefahr von Einengung, Formalisierung und Standardisierung.

Bei der konkreten Ausformung dieses Spannungsverhältnisses in den nationalen Wissenschaftssystemen wird die Spaltung sichtbar zwischen Wissenschaftshandeln – im Sinne eines auf Wissensproduktion ausgerichteten Handelns – und Forschungshandeln als Einbettung dieser Tätigkeit in den lokalen Kontext. Bei einem Vergleich zwischen den europäischen Ländern wird deutlich, wie unterschiedlich die Zeithorizonte sind, im Rahmen derer Handlungen gesetzt werden, wie breit gestreut das Spektrum von Erwartungshaltungen ist, was Wissenschaftspolitik im konkreten bedeutet und vieles mehr. Die Evaluation ist ein Ort, an dem sich solche Unterschiede in besonders deutlicher Weise manifestieren.

Evaluation muß somit im Spannungsfeld zwischen Qualität der Einrichtung, von außen entgegengebrachtem Vertrauen, veränderten Erwartungshaltungen, limitierten und zu legitimierenden Budgets, Verschiebungen in und zwischen Forschungsfeldern und vielem mehr verortet werden. Daraus leiten sich dann auch die zentralen Fragestellungen ab: Welche Aufgaben und Funktionen werden der Evaluation im wissenschaftspolitischen Feld zugeordnet bzw. welche Erwartungen werden von den verschiedenen Seiten an sie gerichtet? Inwieweit kann man mit Evaluation Politik unterstützen oder inwieweit ist nicht Evaluation schon selbst politisches Handeln?

1. Qualität als „bewegliches Ziel“ – Über die Schwierigkeit, Qualität wissenschaftlicher Leistungen „festzumachen“

Gerade in der Diskussion um den Begriff Qualität und dessen Rolle im Wissenschaftssystem wird immer wieder mit Nachdruck darauf verwiesen, daß die Auseinandersetzung darüber immer integraler Bestandteil und Motor wissenschaftlichen Arbeitens gewesen sind. Wir könnten jetzt in der Geschichte zurückgehen und uns damit auseinandersetzen, wie von den ersten Institutionalisierungsschritten der Wissenschaft in Form der wissenschaftlichen Gesellschaften (Ben-David 1991) an bis hinauf zu der von Robert K. Merton (Merton 1985a) postulierten und heftig diskutierten Norm des „organisierten Skeptizismus“ immer wieder eine Frage im Zentrum stand: die nach Standards, die es zu setzen und einzuhalten gilt. Diese Standards trugen/tragen maßgeblich dazu bei, Grenzlinien zu definieren, zu verteidigen oder zu verschieben, und zwar zwischen innen und außen,

also zwischen dem, was als Wissenschaft und dem, was als Nicht-Wissenschaft anzusehen ist, aber auch innerhalb der Wissenschaft zwischen denen, die im Zentrum der Wissenschaftslandschaft angesiedelt sind und jenen an der Peripherie (Gieryn 1995). Wesentlich wäre ebenfalls anzuführen, daß es der wissenschaftlichen Gemeinschaft gelang, sich als geschlossenes Teilsystem der Gesellschaft zu etablieren und trotz der immer wieder stattfindenden Veränderungen im Umfeld die weitgehende Kontrolle über die Produktion des Nachwuchses (also über ihre eigene Erneuerung), über die zentralen zu bearbeitenden Fragestellungen und über die Qualität der erarbeiteten Lösungen zu behalten.[5]

Der US-amerikanische Wissenschaftspolitiker Alvin Weinberg war einer der ersten, der sich explizit mit der Frage wissenschaftlicher Entscheidungskriterien auseinandersetzte und damit gewissermaßen eine seither nicht mehr enden wollende öffentliche Diskussion darüber initiierte. In seinem 1963 publizierten Artikel mit dem Titel „Criteria for Scientific Choice“ (Weinberg 1963)[6] unterschied er explizit zwischen wissenschaftsinternen und wissenschaftsexternen Entscheidungsgrundlagen. Schienen für die traditionelle Kleinforschung jene Kriterien ausreichend, die innerhalb der wissenschaftlichen Gemeinschaft ausgehandelt und zur Anwendung gebracht wurden, so sah man bei Großforschungsprojekten erstmals auch die Notwendigkeit externer Kriterien. Wissenschaftliche Forschung in einem spezifischen Feld sollte nicht nur einen Beitrag zu technologischen Anwendungen und damit zum wirtschaftlichen Wachstum leisten, sondern auch zur Verbesserung der Lebensqualität, der Gesundheit und der Sicherheit beitragen und schließlich auch für andere wissenschaftliche Felder von Bedeutung sein und auf diesem Weg Innovationen hervorbringen (Stichwort: Transdisziplinarität).

Mit diesen Überlegungen markiert Weinberg den Beginn einer wenngleich zum Teil aufgezwungenen Öffnung des Wissenschaftssystems, die sich in den folgenden Jahrzehnten noch akzentuieren würde (Gibbons u. a. 1994). Qualität mußte in einer anderen, erweiterten Weise gedacht werden und die Bedeutung der Wissenschaftspolitik als Verhandlungsraum zwischen Wissenschaft und Gesellschaft wurde immer deutlicher sichtbar. Durch diese Entwicklung wurde auch offensichtlich, daß Qualität nie in einer abstrakten, akteursunabhängigen oder gar zeitlich stabilen Weise festgemacht werden kann. Die in der Literatur unternommenen Versuche, die unterschiedlichen Dimensionen des Qualitätsbegriffes herauszuarbeiten, geben ein klares Zeugnis davon. Einige häufig angeführte Dimensionen sind in Tabelle 1

[5] Interessant hierfür ist die Periode nach dem Zweiten Weltkrieg. Durch das Eindringen von Politik in das wissenschaftliche Feld entstanden neue Strukturen (wie etwa Forschungsförderungseinrichtungen, Forschungsprogramme etc.). Relativ schnell gelang es den Wissenschaftlern, in diesen neuen Strukturen die entscheidenden Positionen einzunehmen und durchzusetzen, daß Entscheidungen nur durch ihre Expertise möglich sind. So wurden Peer Review Verfahren bei der Vergabe von Projekten eingeführt, Expert Panels mit Wissenschaftlern besetzt und vieles mehr.

[6] Siehe auch seinen Aufsatz, in dem er 25 Jahre später die damaligen Einschätzungen kommentiert (Weinberg 1989).

zusammengefaßt (Van Vught 1997), um aufzuzeigen, wie unterschiedlich und dennoch gleichermaßen legitim diese Perspektiven sind. Wenn wir von der Qualität wissenschaftlicher Leistungen sprechen, ist es daher wesentlich, den institutionellen Rahmen zu benennen, in dem sie ausgehandelt wird, offenzulegen, wer die daran beteiligten Akteure sind und explizit zu machen, welche Ziele mit einer Evaluation der Qualität wissenschaftlicher Leistungen verbunden sind.

Tabelle 1

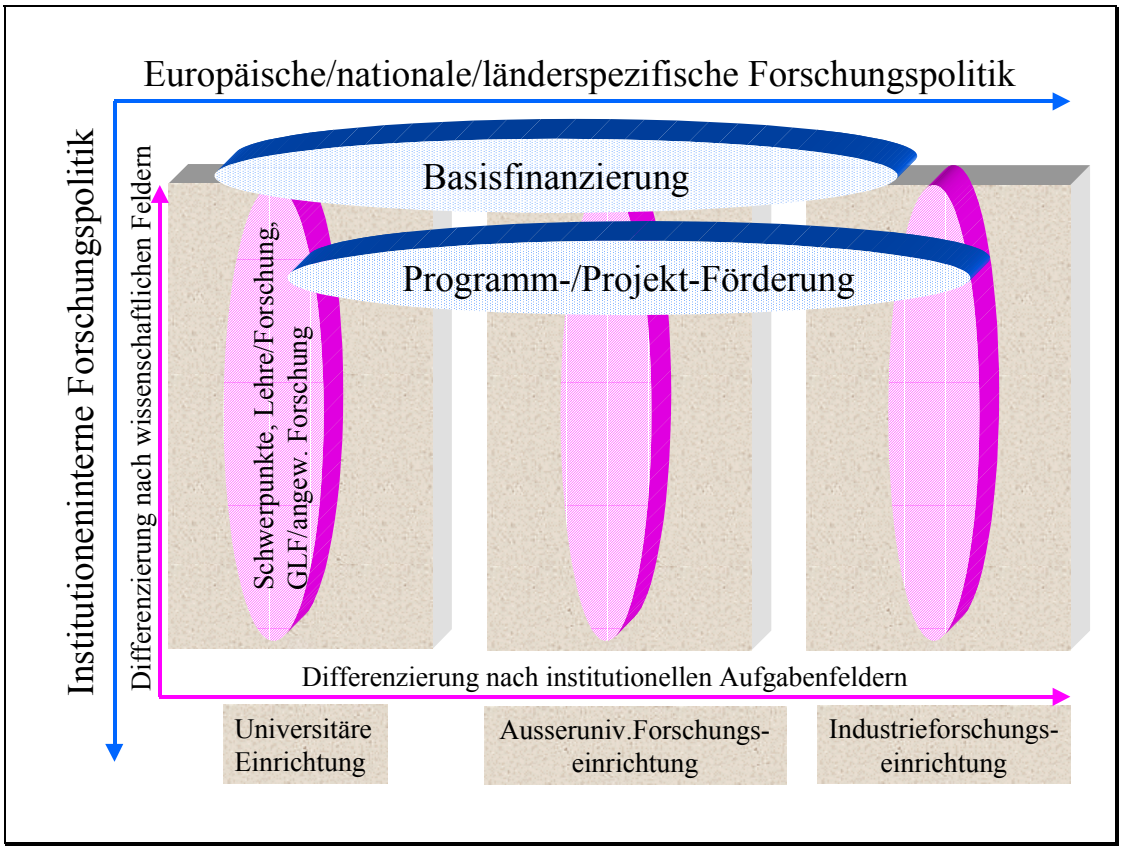
Optionen für eine Definition des Begriffes Qualität in Abhängigkeit von der Anwendung bzw. den Erwartungen an das produzierte Wissen oder die zu erbringenden Dienstleistungen

- Außergewöhnlich innovative wissenschaftliche Leistungen („excellence“) gemessen an von der wissenschaftlichen Gemeinschaft definierten Kriterien
- Verfeinerung bestehender Wissensformen oder Innovation durch eine Transferleistung (entweder zum Anwendungssektor oder zu anderen Disziplinen)
- Passgenauigkeit mit einem gesetzten Ziel (dies kann sowohl von innen als auch von außen vorgegeben werden)
- Fehlerfreiheit bei wachsender Präzision
- Auswirkungen auf weitere Arbeiten („impact“)
- Stimmigkeit im Verhältnis von Aufwand und erbrachter Leistung
- Nützlichkeit für andere Teilsysteme der Gesellschaft (Wirtschaft, Ausbildung, ...)
- etc.

Die aufgezeigten Probleme mit der Qualitätsdiskussion verweisen aber auch auf die Verwobenheit des Wissenschaftssystem und die zum Teil völlig unterschiedlichen zu evaluierenden Bereiche. In Abbildung 1 wurde versucht, dieses komplexe und von sehr unterschiedlichen Organisations- und Aufgabenstrukturen durchzogene wissenschaftliche Feld graphisch darzustellen. Zum einen stehen wir einer nach nationalen Kontexten sehr unterschiedlichen institutionellen Ausdifferenzierung gegenüber. In der Graphik wurden universitäre Einrichtungen von außeruniversitären und diese wiederum von Forschungseinrichtungen der Industrie unterschieden. Jede von ihnen hat unterschiedliche Aufgaben und entsprechende Strukturen, um diese zu erfüllen. Innerhalb dieser Institutionen stehen wir einer nochmaligen Differenzierung nach wissenschaftlichen Feldern gegenüber, die jeweils durch eigene Rituale, Entscheidungsmuster, Publikationsverhalten, Beurteilungskriterien usw. charakterisiert

sind. Diesen Unterscheidungen überlagert sind nochmals zwei Einflußebenen zu berücksichtigen. Zum einen ist dies die europäische, nationale bzw. länderspezifische Forschungspolitik, welche versucht, die jeweiligen Interessen über entsprechende Programm- und Projektförderung, aber zum Teil auch schon über die Basisfinanzierung zu implementieren. Hierbei kann es sowohl zu Verstärkungseffekten[7] zwischen den drei angesprochenen Ebenen kommen, aber durchaus auch zu Spannungsverhältnissen. Zum anderen versuchen die wissenschaftlichen Institutionen selbst unter wachsendem Wettbewerbs- und Rechtfertigungsdruck ein ihnen eigenes und von den anderen unterschiedliches Profil zu entwickeln und es durch eine „hausinterne" Forschungspolitik auch umzusetzen und zu erhalten. Auf das sich in steter Entwicklung befindliche wissenschaftliche Feld wirken also gleichzeitig sehr unterschiedliche Kräfte, die sich ihrerseits verändern: Qualität wird zu einem beweglichen Ziel.

Abbildung 1: Verschiedene Kräfte, die im Wissenschaftssystem wirksam sind



[7] Dies läßt sich bei EU-Programmen durchaus beobachten. Wenn ein Bereich von der EU als strategisch wesentlich angesehen wird und finanzielle Mittel in diesen fließen, erweist es sich zumeist auch als einfacher, auf nationaler Ebene Zusatzfinanzierungen für diesen Bereich zu erhalten.

Ausgehend von diesen verschiedenen und nur lose gekoppelten Wirkungskräften, die das wissenschaftliche Feld durchdringen, entstehen bisweilen gravierende Spannungen. Diese ergeben sich vor allem aus einer mangelnden Festlegung der relativen Bedeutung der verschiedenen Einflüsse in einem definierten institutionellen Raum. So sind vielfach weder die Hierarchien zwischen den unterschiedlichen Aufgaben noch innerhalb der Aufgabenfelder klar vorgegeben. Die Universitäten sind ein sehr gutes Beispiel hierfür. Kann man sich noch auf eine Aufzählung der Aufgaben einigen (Grundlagen- und angewandte Forschung, Vor-, Aus- und Weiterbildung, Erhaltung von Wissen, ...), so sieht es mit der Gewichtung zwischen diesen Aufgaben schon sehr viel schwieriger aus. Dadurch entstehen eine ganze Reihe von Konflikten, die gerade bei Evaluationen zu Tage treten. So etwa zwischen dem Individuum mit seinen/ihren Karriere- und Entwicklungsmöglichkeiten und der Institution, welche bisweilen andere Prioritäten und kurzfristige Ziele hat, zwischen unterschiedlichen Aufgabenschwerpunkten wie Forschung und Lehre, zwischen Außenanforderungen und Binnenentwicklungen, zwischen Grundlagenforschung und anwendungsbezogener Erkenntnisproduktion und vieles mehr. Konkret manifestiert sich diese Problematik in sehr unterschiedlicher Weise, je nach nationalen/lokalen Strukturen und den dort herrschenden Bedingungen, aber auch je nachdem, wo Evaluationen ansetzen und auf welche Aspekte dabei fokussiert wird.

Das eben Gesagte macht deutlich, daß es für einen sinnvollen Einsatz von Evaluationen wesentlich ist,

- sich genauer, als dies vielfach der Fall ist, mit dem Ziel einer Evaluation auseinanderzusetzen,
- ein besseres Verständnis für die in einem spezifischen Teilbereich beteiligten Akteure zu entwickeln und
- die Idee einer einfachen Vergleichbarkeit zwischen Bereichen, Ländern und Institutionen aufzugeben.

Evaluationen können bestenfalls die unterschiedlichen Positionierungen der bewerteten Einheiten im Detail ausloten, Vergleiche zwischen einzelnen Facetten möglich machen und schließlich die Basis für wissenschaftspolitische Entscheidungen bilden.

2. Zur zentralen Rolle der Zieldefinition bei Evaluationen

In der mittlerweile sehr umfangreichen Literatur zum Thema Forschungsevaluation herrscht Einigkeit über die Bedeutung der Zieldefinition als erster wesentlicher Schritt im Rahmen einer Evaluation. Sie soll für alle Beteiligten und Betroffenen Klarheit darüber schaffen, warum eine Leistung festgestellt werden soll, welche Erwartungen an dieses Prozedere gekoppelt sind und wo und wie die gewonnenen Erkenntnisse zur Umsetzung gelangen sollen.

Trotzdem dies weitgehend bewußt ist, entstehen sehr viele Probleme im Bereich der Evaluation genau durch ein Nichtbeachten der zuvor definierten Ziele und der damit auferlegten Grenzen. Um dies besser zu veranschaulichen, kann man eine vom US-amerikanischen Bildungsforscher Martin Trow (Trow 1994) entwickelte Typologie von Bewertungsmechanismen heranziehen. Zwei grundlegende Parameter spannen dabei eine Matrix von vier Möglichkeiten auf: „Funktion der Evaluation", welche noch in „unterstützend" und „bewertend" differenziert wird, und „Ausgangspunkt der Bewertung", wobei „intern" und „extern" unterschieden werden.

Abbildung 2: Typologie der Forschungsevaluationen nach M. Trow

		Funktion der Evaluation	
		unterstützend	bewertend
Ausgangspunkt der Evaluation	intern	Typus I	Typus II
	extern	Typus III	Typus IV

Die daraus resultierenden vier Typen von Evaluationen sind dann wie folgt charakterisiert:

• Typus I: *Interne unterstützende Evaluation*

Im akademischen Bereich ist dies nach Trow die häufigste Form von Evaluation. Von Experten durchgeführt, die der zu bewertenden Einheit nahestehen und somit über ausreichendes Detailwissen verfügen, erlaubt sie eine Früherkennung von Fehlentwicklungen. Intensive kritische Auseinandersetzungen mit dem eigenen Bereich sind dafür notwendig. Offensichtliche Nachteile wie etwa das Entstehen von Spannungen bzw. das Vermeiden von Konflikten sind allerdings nicht zu übersehen, können zum Teil aber durch das Hinzuziehen von externen GutachterInnen aufgewogen werden. Dabei sollte aber nicht übersehen werden, daß solche regelmäßigen Qualitätssicherungsmechanismen sehr zeitaufwendig sind, ein hohes Maß an Verantwortungsbewußtsein der Beteiligten fordern und von Seiten der Institution die Vorkehrungen getroffen werden müssen, um eine Umsetzung der Empfehlungen auch sicherzustellen.

• Typus II: *Interne bewertende Evaluation*

Dieser Evaluationstypus wird häufig zur direkten ad hoc Entscheidungsfindung etwa im Falle notwendiger Prioritätensetzung oder als Instrument des Krisenmanagements bei Budget- oder Postenkürzungen eingesetzt. In diesem Fall steht also nicht die Kontinuität und die Verbesserung der jeweiligen Einheit im Vordergrund, sondern die Notwendigkeit einer Entscheidung. In zahlreichen Fällen läßt sich hier beobachten, daß hier dann Ergebnisse einer vorangegangenen Typus I–Evaluation einfließen und somit gewissermaßen mißbräuchlich verwendet werden.

• Typus III: *Externe unterstützende Evaluation*

Diese Form der Evaluation, die einen wesentlichen Beitrag zur Optimierung wissenschaftlicher Leistungen einer Einheit leisten könnte, findet dennoch relativ selten statt. Da externe Experten eingesetzt werden, besitzen solche Evaluationen meist größere Glaubwürdigkeit und Akzeptanz sowohl innerhalb der wissenschaftlichen Gemeinschaft wie auch im wissenschaftspolitischen Umfeld und der breiteren Öffentlichkeit. Ihre Kostenaufwendigkeit und Organisationsintensität machen sie aber als rein unterstützendes Instrument weniger attraktiv.

• Typus IV: *Externe bewertende Evaluationen*

Solche etwa von Ministerien in Auftrag gegebene Evaluationen sind vielfach auf Aspekte von Management und Kontrolle oder auf anstehende Entscheidungen ausgerichtet. Dadurch werden Druck, Spannung und zum Teil Mißtrauen erzeugt, was in der Folge dazu führt, daß eher der Präsentationscharakter als die tatsächliche wissenschaftliche Leistung in den Vordergrund tritt. Es wird alles darauf ausgerichtet, zu überzeugen, und damit wird man der Möglichkeit verlustig, offen über Probleme und Schritte zu deren Behebung zu verhandeln.

Was sich durch eine solche Kategorisierung erkennen läßt, ist die Notwendigkeit einer sehr klaren Positionierung von Evaluationen in einem der vier Felder. Denn die verwendeten Methoden und Prozedere funktionieren und greifen nur in diesem Kontext. Grenzüberschreitungen schaffen nicht nur technische Probleme, sondern können einen massiven Vertrauensbruch mit sich bringen und so die Optionen einer Verbesserung wieder zunichte machen.

Schließlich könnte man hier in Klammern noch anfügen, daß die Ergebnisse von Evaluationen im Grunde oft nicht wirklich überraschende Ergebnisse liefern und von Kennern der Situation auch informell eingeschätzt hätten werden können. Der Wert der Evaluation liegt vielmehr im Prozeß selbst, im Sichtbar- und Nachvollziehbarmachen sowie im formellen Darstellen von gewonnenen Erkenntnissen. Erst durch diesen Akt der Externalisierung werden Facetten und Positionen hinterfragbar und damit bisweilen auch veränderbar.

3. Wer sind die beteiligten Akteure?

Die Bedeutung der an der Qualitätsdefinition beteiligten Akteure hat Maurice Goldsmith in seinem Buch „The Science Critic“ klar auf den Punkt gebracht: „Die Auseinandersetzung dreht sich nicht länger darum, ob es eine öffentliche Teilnahme und Kontrolle von Wissenschaft geben wird, sondern wer an der Einrichtung von Kontrollen beteiligt sein wird, wie solche Kontrollen organisiert werden und wie diese detaillierten Entscheidungen die Natur und die Forschungsprozesse beeinflussen werden“ (Goldsmith 1986).

Es geht also darum, zu verstehen, wer genau an diesem Definitionsprozeß beteiligt ist, in welcher Form solche Qualitätskontrollen organisiert sind und welche Erwartungen und Werte hier von außen in das Wissenschaftssystem eingebracht werden. Selbst wenn nur Mitglieder der wissenschaftlichen Gemeinschaft als Akteure im Evaluationsprozeß zugelassen sind, hat etwa Robert Merton in seinen frühen wissenschaftssoziologischen Arbeiten bereits auf bestimmte Auswahl- und Verstärkungsphänomene verwiesen, die er unter dem Begriff Matthäus-Effekt (Merton 1985b)[8] zusammengefasst hat. Dies bedeutet, daß bestehende Machtverhältnisse, soziale Relationen, Hierarchien etc. durch die Strukturen des Wissenschaftssystems ständig weiter akzentuiert werden. Auch wenn versucht wird, diesem Phänomen durch reflexives Verhalten entgegenzuwirken, ist von der Entscheidung über die Gutachter bis hin zur Auswahl der zu evaluierenden Einheit und deren Miteinbeziehen in die Auswahl der zu bewertenden Parameter und zu erhebenden Daten alles bereits mitbestimmend für mögliche Ergebnisse einer Evaluation.

Darüber hinaus haben rezente Arbeiten, wie etwa die von Gibbons und anderen (Gibbons u. a. 1994) auch auf eine Erweiterung der Gruppe jener hingewiesen, die legitimerweise (wenngleich zum Teil indirekt) an der Qualitätsdefinition beteiligt sind. Die Tatsache, daß wissenschaftliches Wissen insgesamt verstärkt im Kontext der Anwendung produziert und gesehen wird und gleichzeitig von öffentlicher Seite erhöhter Rechtfertigungsdruck auf das Wissenschaftssystem ausgeübt wird, fließt automatisch auch in die Bewertungen von Leistungen ein. Es ist also einerseits durch Ausdifferenzierung und Spezialisierung dem Wissenschaftssystem gelungen, sich immer deutlicher von seinem Umfeld abzugrenzen, dies hat aber gleichzeitig dazu geführt, daß neue Formen der Wechselwirkung mit dem Umfeld entstanden sind. Darüber hinaus werden immer öfter Fragestellungen von außen an das Wissenschaftssystem herangetragen und diese lassen sich vielfach nicht innerhalb der relativ rigiden disziplinären Grenzen einordnen. Die Qualität von in diesem Zusammenhang erbrachten wissenschaftlichen Leistungen ist somit auch nicht mehr alleinig aufgrund interner Kriterien beurteilbar.

[8] Für eine Diskussion der zusätzlichen geschlechtsspezifischen Mechanismen, die zum Tragen kommen, siehe Rossiter (1993).

Das eben Gesagte kann am Beispiel der Universitäten sehr deutlich illustriert werden: Es stellt sich hier die Frage der an der Qualitätsdiskussion zu beteiligenden Akteure gleich in zweierlei Weise, nämlich in Lehre und Forschung. Bei der Lehre fragt es sich, ob es nun die Institution Universität selbst, die Studierenden als „Konsumenten" dieser Lehrleistungen, Vertreter des Arbeitsmarkts, die Absolventen beschäftigen, oder Financiers der universitären Lehre (etwa Ministerien) sind, die bei einer Bewertung maßgeblich beteiligt sein sollten. Hier könnten also als Ergebnis einer solchen Evaluation Forderungen für eine Verbesserung formuliert werden. Auf der anderen Seite hat aber die Universität ebenso wesentliche Aufgaben in der Forschung. Und auch hier stellt sich die Frage, ob es nur FachkollegInnen sein sollten, die die Leistung bewerten, oder ob nicht darüber hinaus einerseits Personen zugelassen werden sollten, die sowohl die Angemessenheit der Forschungsorganisationsstrukturen, Optionen der Personalentwicklung etc. beurteilen (können) und andererseits Vertreter naher Disziplinen, die dann auch die Relevanz und das Entwicklungspotential aus einer entsprechenden Distanz beurteilen können.

Wie auch immer in den beiden Bereichen – Lehre und Forschung – die Entscheidung ausfällt, so löst sich jedoch keinesfalls das Problem der relativen Wichtigkeit der beiden Bereiche innerhalb der Institution Universität – und diese Einschätzung ist wohl je nach Position und Sichtweise sehr divergierend.

4. Zur Vielzahl aktueller Praktiken in Europa – Evaluation als nationale/lokale Kultur

Während eine umfangreiche Literatur der Wissenschaftsforschung auf den Einfluß lokaler, kultureller bzw. nationaler Kontingenzen in der Entwicklung des Wissenschaftssystems verweist, so gilt selbiges auch für die Auswahl der Evaluationsmethoden und die implementierten Abläufe. Es gibt zwar ein international anerkanntes Set von Methoden der Leistungsbewertung von *Peer Review* Verfahren der verschiedensten Art bis hin zu unterschiedlichen Formen quantitativer Indikatoren, aber gleichzeitig ist es gerade die spezifische Auswahl und die Art und Weise, wie diese zum Einsatz kommen, welche ausschlaggebend für die konkrete Evaluation sind. Interessant ist, in diesem Zusammenhang anzumerken, daß es zwar eine durchaus differenzierte Methodenkritik gibt, diese aber gleichzeitig immer wieder ignoriert wird. So wurde bei Peer Review-Evaluationen etwa immer wieder auf das Phänomen der „Old Boys Networks"[9] verwiesen und die Einflüsse von männlich besetzten und gut verankerten Netzwerken sichtbar gemacht. Ein Beispiel dafür ist der in der Zeitschrift *Nature* publizierte Fall des Swedish Medical Research Council, welches – so die Autorinnen – bei der Antragsbegutachtung bei gleicher Leistung der AntragstellerInnen Anträge von Frauen wesentlich häufiger ablehnte als die männ-

[9] Für die Kritik am Peer Review System siehe Travis und Collins (1991), Chubin und Hackett (1990).

licher Kollegen.[10] Darüber hinaus könnten noch eine ganze Reihe anderer Phänomene angeführt werden, etwa eine verstärkte Anpassung an *Mainstream* Forschung oder die Schwierigkeiten, mit denen interdisziplinäre Forschungsprojekte konfrontiert sind, und vieles mehr.

Aber auch die Verfahren, die sich stark an quantitativen Indikatoren orientieren, waren einer heftigen Kritik ausgesetzt, da sowohl die dabei verwendeten Konzepte, aber auch die Kontexte, in denen sie schließlich eingesetzt wurden, hochgradig unscharf definiert waren. Waren Peer Review Verfahren zu Anfang der Institutionalisierung der Wissenschaft entwickelt worden und daher ein untrennbar mit der Entwicklung des Wissenschaftssystems verbundenes Vorgehen, so entstanden Wissenschaftsindikatoren erst durch die enge Verquickung mit Wissenschaftspolitik (also etwa ab den 70er Jahren), für die sie entscheidungs- und steuerungsrelevante Informationen liefern sollten. Indikatoren versuchen dabei Realitäten im Wissenschaftssystem abzubilden, wobei sie aber gleichzeitig durch ihre Existenz und ihre Verwendung diese Realitäten miterzeugen. Vor allem aber sind Indikatoren in einem bestimmten Kontext entstanden, der in ihnen verankert ist, den sie aber zumeist im Laufe ihrer Anwendung „vergessen". Durch diese Dekontextualisierung erhalten sie eine scheinbare Robustheit, die vielfach die Attraktivität dieser Methode ausmacht (sie wirkt „objektiv" und nachvollziehbar), aber auch die eigentliche Gefahr darstellt. Auch hier werden im Grunde eher Mainstream Phänomene gegenüber radikalen (zum Teil interdisziplinären) Innovationen bevorzugt, der Prozeßaspekt von Forschung wird kaum berücksichtigt, und im Grunde verleitet ein verstärkter Einsatz von Indikatoren zu meist rigiden Handlungsschemata von Seiten der Forscher.[11]

Die Methodendiskussion aufzuarbeiten, die auch kontinuierlich zu Verfeinerungen der Methoden und Verfahren geführt hat, wäre wesentlich zu umfangreich für einen zeitlich beschränkten Rahmen, ebenso wie ein wirklicher Vergleich zwischen den verschiedenen nationalen Systeme. Dennoch möchte ich nun anhand von einigen beispielhaft ausgewählten Perspektiven[12] aufzeigen, bis zu welchem Grad die verschiedenen aufgezeigten Problembereiche in den nationalen Kontexten ganz unterschiedlichen Lösungen zugeführt werden. Dabei sind die jeweiligen Lösungsansätze

[10] Als konkretes Beispiel für den Einfluss von Peers auf die Qualitätsdefinition siehe die Fallstudie von Wenneras und Wold (1997).

[11] Literatur zum Thema Forschungsindikatoren: Weingart, Sehringer und Winterhager (1991); Weingart und Winterhager (1984). – Für eine kritische Diskussion der Anwendung von Forschungsindikatoren siehe z. B. Edge (1979), Weingart, Sehringer und Winterhager (1991), Woolgar (1991), Hornbostel (1997).

[12] Bei der Zusammenstellung dieser ausgewählten Perspektiven wurde auf umfangreiches Material zurückgegriffen, welches im Rahmen eines Projekts „Evaluationmaßnahmen im Bereich der Forschung an der Universität Wien" zusammengetragen wurde. Die Quellen sind hier daher nicht im einzelnen angeführt. – Für einen ausgewählten internationalen Vergleich der Evaluationsverfahren siehe auch Felderer und Campbell (1998).

insbesondere im Zusammenhang mit ganz unterschiedlichen Karrieremustern, Personalentwicklungsstrategien, Förderungsmechanismen und schließlich auch mit der institutionellen Struktur des Wissenschaftssystems zu sehen.

Zum ersten ist es von großem Interesse, die *Kopplungsmechanismen* von Evaluation und wissenschaftspolitischen Maßnahmen zu betrachten. Mit der Ausnahme von Großbritannien, wo regelmäßige Evaluationen auch die Basisfinanzierungen wissenschaftlicher Institutionen steuern, ist eine solch rigide Kopplung in Europa unüblich. Gerade der britische Fall wird aber immer wieder als Beispiel dafür angeführt, daß eine zu zentrale Bedeutung von festen Eckdaten im Grunde zu einer strategischen Anpassung der Forscher an das System führt und so bisweilen die intendierten Veränderungen nicht erreicht werden. In den anderen Ländern sind die Auswirkungen von Evaluationen zumeist eher indirekter Natur oder beschränken sich nur auf bestimmte Facetten des Systems, da sie immer Ergebnisse von zusätzlichen nach der eigentlichen Evaluation stattfindenden Aushandlunsgprozessen sind.

Auch bei der Frage nach dem *„wie"* einer solchen Qualitätsfeststellung gibt es ganz wesentliche Unterschiede. Dabei geht es nicht nur darum, in den einzelnen Aufgabenkategorien des Wissenschaftssystem entsprechend operationalisierbare Parameter festzulegen, sondern vor allem auch um die schwierigen Fragen der Gewichtung zwischen den verschiedenen Bereichen.

Hier möchte ich nur zwei Beispiele herausgreifen, um die Spanne der unterschiedlichen Behandlung aufzuzeigen. So werden etwa in den Niederlanden (entwickelt auf nationaler Ebene durch die VSNU, der Vereinigung Niederländischer Universitäten) vier Qualitätsdimensionen unterschieden, die wiederum jeweils auf einer fünfteiligen Skala bewertet werden. Diese Dimensionen sind: 1. Qualität im klassischen wissenschaftsinternen Sinn, 2. Relevanz (im Sinne der Aufnahme und Verbreitung dieses Wissens), 3. Produktivität, also die Frage nach der Relation von Input und Output, sowie 4. Harmonie mit den gesetzten Zielen, Potential für die Zukunft. Die Relation der vier Parameter ist allerdings nicht festgelegt, und daher ist es leicht vorstellbar, daß sich hier selbst bei einer relativen Einigkeit über Ergebnisse in den einzelnen Kategorien ein großer Verhandlungsraum auftut.

In Großbritannien hingegen wird im Rahmen der „Research Assessment Exercises" die Beurteilung der Qualität in Form einer Note auf einer siebenteiligen Notenskala zum Ausdruck gebracht, was aus meiner Sicht zu einer frühen Reduktion der Komplexität führt. Aushandlungsprozesse müßten also eine Ebene davor – im Vergleich zum niederländischen Modell – stattfinden, nämlich bereits bei der aus den Einzelperspektiven resultierenden Festlegung der Gesamtbeurteilung. Diese ist dann von großer Wichtigkeit, da diese Zahl in eine Formel eingeht, die ihrerseits wiederum die Basisfinanzierung festschreibt.

Dies bringt mich zum dritten wesentlichen Unterschied in den Evaluierungskulturen, nämlich zur Frage nach den *Einheiten, die einer Bewertung unterzogen werden*. Hier zeigt sich sehr deutlich – wie bereits erwähnt – die enge Kopplung von Evaluation mit den nationalen Forschungsstrukturen. Während in den Niederlanden im Grunde die sogenannten Forschungsprogramme[13] (jedes Institut ist in mehreren Forschungsprogrammen involviert; die evaluierte Einheit ist nicht deckungsgleich mit der Organisationseinheit) einer Bewertung unterzogen werden, sind es in Großbritannien die *Units of Assessment* (ähnlich einer Einteilung in Disziplinen), im Rahmen des französischen CNRS fachgruppenähnliche Einheiten oder waren es in Österreich bislang Disziplinen (oder Teile von diesen). Hier stellen sich gleich zwei Fragen: die erste nach dem institutionellen Rahmen, der es nach der Evaluation ermöglichen würde, auch positive Auswirkung auf Systemebene zu erreichen. Oder expliziter ausgedrückt: Sind es tatsächlich die bewerteten Einheiten, die dann auch Veränderungen durchsetzen können? Die zweite Frage berührt das Problem der tatsächlichen Vergleichbarkeit von Leistungen (etwa zwischen Forschungsprogrammen oder Units of Assessment) und die damit in Zusammenhang stehenden Entscheidungen über die Gewichtungen in der respektiven Unterstützung nach einer solchen Evaluation.

Ein vierter interessanter Aspekt, der große nationale Variationen aufweist, ist die *Transparenz solcher Verfahren und die Publikation der Ergebnisse*. Auch hier könnte man beispielhaft sagen, daß etwa in Großbritannien sowohl die zur Anwendung gelangende Bewertungsmethode als auch das genaue Prozedere im Vorhinein publiziert werden. Auch die Ergebnisse werden in Form eines Rankings öffentlich zugänglich gemacht. In den Niederlanden wurde 1994 ein sogenanntes Protokoll publiziert, in dem alle Grundzüge der Evaluationen festgeschrieben sind und welches sowohl den Peers als auch den beurteilten Wissenschaftlern als Rahmen dienen soll. Die Veröffentlichung der Ergebnisse ist weit weniger standardisiert als in Großbritannien.

Die letzten wesentlichen Unterschiede liegen in der *Methoden- und Gutachterwahl*. So sehen die Niederlande im Rahmen der Forschungsevaluation immer sowohl Selbst- als auch Fremdevaluation vor, was eine extensive Auseinandersetzung der Wissenschaftler mit der eigenen Leistung zur Folge hat. Darüber hinaus setzen sie stark auf externe Gutachter. In Frankreich wird etwa im Rahmen der Entscheidungen von CNRS Forschungsschwerpunkten fast ausschließlich auf nationale Gutachter (wenn überhaupt) rekurriert, wobei die angewandten Kriterien nicht wirklich explizit gemacht werden. In den bisher in Österreich durchgeführten Evaluationen hatte man bei der Gutachterwahl sowohl auf deren Internationalität, aber immer auch auf einen hohen Grad der Akzeptanz der Gutachter durch die wissenschaftliche Gemeinschaft geachtet (bei der Evaluation der Physik wurden die Gutachter von der Österreichischen Physikalischen Gesellschaft ausgewählt).

[13] Es wird zwar auch eine Einteilung nach Disziplinen vorgenommen, aber evaluiert werden dann die Forschungsprogramme.

Was diese fünf nur sehr kurz angesprochenen Punkte aufzeigen, ist ihre Wichtigkeit für ein Gelingen einer Evaluation, wobei ich hier unter Gelingen von der wissenschaftlichen Gemeinschaft auch als positiv akzeptierte Veränderungen verstehe. In jedem Kontext müssen diese Punkte spezifisch ausgehandelt und weitgehender Konsens zwischen möglichst vielen Beteiligten und Betroffenen hergestellt werden.

5. Die Forderung nach einer Bewertung der Bewertung

Bei aller Evaluationseuphorie – Evaluation wurde zum Allheilmittel hochstilisiert –, die sich in den letzten Jahren auch in politischen Kreisen breit gemacht hat und zum Teil auch die wissenschaftlichen Institutionen erfaßt hat, darf man zumindest zwei grundlegende Dinge nicht aus den Augen verlieren. Wenn man das bisher Gesagte ernst nimmt, nämlich, daß sich sowohl das Wissenschaftssystem als auch der Qualitätsbegriff in einem stetigen Wandlungsprozeß befinden, so muß dies auch für die zum Einsatz kommenden Instrumentarien Gültigkeit besitzen. In regelmäßigen Abständen müssten also nicht nur die Forschungs- oder Lehrsituation und die dazugehörenden Ergebnisse bewertet, sondern auch ein genauerer Blick darauf geworfen werden, ob die angewendeten Instrumentarien noch der jeweils aktuellen Situation angemessen sind und ob Evaluation auch die gewünschten Veränderungen bewirkt hat.[14]

Dies bringt uns zum zweiten Punkt, nämlich der Tatsache, daß Evaluationen nicht nur positive Auswirkungen für wissenschaftliche Institutionen bringen können, sondern auch relativ hohe Kosten verursachen. Vereinfacht gesagt: Es ist sinnvoll und notwendig, in regelmäßigen Abständen eine Kosten-Nutzen-Rechnung anzustellen. Bei den Kosten ist zum einen der explizite Aufwand abzuschätzen, z. B. Personal für die Organisation, Gutachterkosten etc., zum anderen aber auch implizite Aufwendungen wie das Zeitbudget des Personals, welches durch zu häufige Evaluationen durchaus belastet werden würde. Solche Kosten-Nutzen-Einschätzungen können allerdings sehr schwierig werden, da vor allem bei größeren Forschungseinheiten oft nicht einfach abschätzbar ist, in welcher Weise, an welcher Stelle im System und mit welcher Intensität von außen initiierte Veränderungen auch tatsächlich ihre Wirkung zeigen werden.

6. Evaluation als Aushandlungs- und Veränderungsprozeß

Der immer wieder angesprochene Prozeßcharakter der Qualitätssicherungsmaßnahmen kommt nun konkret in zumindest dreierlei Weise zum Tragen.

[14] In Quebec etwa wird in regelmäßigen Abständen eine Evaluierung der Lehrevaluierungen an Universitäten durchgeführt, um die Zweckmäßigkeit dieser Maßnahmen sicherzustellen.

Zum ersten kann die Veränderung und Optimierung der Leistungen einer wissenschaftlichen Institution nur als kontinuierlicher, die verschiedenen Aufgaben und Facetten berücksichtigender Prozeß wirklich erreicht werden. Das bedeutet, daß nicht auf allen Ebenen gleichzeitig mit Evaluationen begonnen werden kann, sondern zeitlich verschobene Anlaufphasen einzuplanen sind, damit die notwendigen Diskussionen und Aushandlungsprozesse auch wirklich stattfinden und Veränderung möglich wird.

Zweitens muß der Prozeßcharakter auch auf der Ebene der Methoden und Instrumentarien seine Verankerung finden. Es muß nicht nur eine Palette von unterschiedlichen, den zu evaluierenden Bereichen angepassten Methoden entwickelt werden, sondern diese sollten auch mit Hilfe sukzessiver Testverfahren an die Spezifitäten der jeweiligen Forschungseinrichtung angepasst werden. Im Methodenbereich besteht zwar die Möglichkeit, auf eine Fülle von internationalen Erfahrungen zurückzugreifen, diese können aber nicht direkt und ohne Adaptierung an den lokalen Kontext übernommen werden. Besondere Aufmerksamkeit sollte dem Phänomen geschenkt werden, daß in Phasen großen Rechtfertigungsdrucks und Unsicherheit die Tendenz besteht, auf starre Indikatorensysteme und eher quantitativ orientierte Methoden zurückzugreifen. Diese eignen sich zwar als eine ergänzende Basis oder als ein Überblicksinstrumentarium, beinhalten jedoch – wie bereits erwähnt – die Gefahr, eher systemverstärkend/-konservierend als systemerneuernd zu wirken. Bei der Schaffung und Ausgestaltung aller Instrumentarien geht es um eine frühzeitige Auseinandersetzung mit den unterschiedlichen beteiligten und betroffenen Akteuren, aber auch um ein regelmäßiges Hinterfragen der zur Leistungsfeststellung eingesetzten Mittel.

Und schließlich sollte darauf verwiesen werden, daß die erhoffte Optimierung sicherlich nicht schlagartig zu erwarten ist, sondern nur langsam greifen und damit sichtbar werden wird. Es geht ja um einen recht komplexen Prozeß der Aushandlung zwischen den „von außen“ herangetragenen Wünschen, Beschwerden und Anregungen und den „innen“ entwickelten Vorstellungen.

7. Die Klammer zwischen Forschungspolitik und Evaluation

Sechs kurze thesenhafte Anmerkungen möchte ich an das Ende meiner Ausführungen stellen und so auch die Verbindung zwischen der Qualitätsdiskussion und Forschungspolitik knüpfen.

- Es macht keinen Sinn – und hier möchte ich mich den Überlegungen von Martin Trow anschließen – immer öfter, mit immer neuen Qualitätsfeststellungsverfahren Teile des Wissenschaftssystems zu untersuchen, ohne vorher verstanden zu haben, warum zum Teil bereits bestehende/angewandte Verfahren nicht die gewünschten Ziele erreicht haben. Evaluationen sind in komplexe soziale Prozesse eingebettet, und diese gilt es in erster Linie zu verstehen, um

eine Optimierung bewirken zu können. Wissenschaftspolitische Maßnahmen können daher nicht nur auf der Makroebene ansetzen, sondern müssen eine Passform mit den mikro-soziologischen Strukturen anstreben.

- Forschungspolitik muß auch bedeuten können, Bereiche zu fördern und auszubauen, die bislang nicht die gewünschten Qualitätsstandards erreicht haben. Es geht also um bewußte Entscheidungen und Weichenstellungen. Denn schlechte Bewertungsergebnisse können ihren Grund auch in mangelnder struktureller und ressourcenmäßiger Ausstattung haben. Evaluation sollte somit keineswegs Politik ersetzen, auch wenn sie natürlich immer schon politische Dimensionen in sich trägt.

- Forschung ist ein dynamischer, vielschichtiger Prozeß, und dies sollte auch in den Evaluationsmethoden und in den Prozedere seinen Niederschlag finden. Einmalige Evaluationen, die ausschließlich dem Setzen von Maßnahmen dienen, sind Momentaufnahmen des Wissenschaftssystems, die sich als trügerisch erweisen könnten. Im Grunde sollte es also darum gehen, Planung und Evaluation stärker miteinander zu koppeln und Qualität vielmehr am Erreichen vorab konsensual erarbeiteter Ziele zu messen.

- Rigide Kopplung von Qualitätsbewertung an Finanzierung erzeugt Anpassung statt Innovation und kann das Wissenschaftssystem nachhaltig schädigen. Wissenschaftliche Entwicklungen haben immer auf einer gewissen Risikobereitschaft der Geldgeber aufgebaut und sind immer vorerst „Versprechen in die Zukunft“. Es gilt also unter den geänderten Rahmenbedingungen eine neue Form der Vertrauensbasis zwischen Wissenschaft und ihren Unterstützern aufzubauen.

- Evaluation ist nicht Ersatz für eine Auseinandersetzung und für Aushandlungsprozesse, sondern – ganz im Gegenteil – sie soll diese fördern, strukturieren und die oft impliziten Kriterien sichtbar und somit hinterfragbar und verhandelbar machen.

- Input-Output Relationen stehen viel zu häufig bei Evaluationen im Vordergrund. Die wissenschaftlichen Institutionen und die darin ablaufenden Prozesse verkommen dabei zu „black boxes“. Dabei sollten gerade diese Transformationsprozesse von Input zu Output im Zentrum stehen, besser verstanden und optimiert werden. Hier liegt das eigentliche Innovationspotential der Evaluation: gleichzeitig Bewußtsein zu schaffen sowohl über die vielschichtigen Rollen und zentralen Funktionen von Qualität im Rahmen wissenschaftlicher Institutionen als auch über die Schwierigkeiten ihrer Definition und Operationalisierung.

Literaturverzeichnis

Bazerman, C. (1988), Literate acts and the emergent social structure of science, in: ders., Shaping Written Knowledge, Madison, Wis. (u.a.), S. 128-150

Ben-David, J. (1991), Scientific Growth. Essays on the Social Organization and Ethos of Science, Berkeley

Cozzens, S. E., P. Healey, A. Rip und J. Ziman (Hg.) (1990), The Research System in Transition, Dordrecht

Chubin, D. E. und E. J. Hackett (1990), Peerless Science, Peer Review and US Science Policy, New York

Edge, D. (1979), Quantitative measures of communication in science: a critical review, in: History of Science XVII, S. 102 -127

Felderer B. und D. F. J. Campbell (1998), Die Evaluation der akademischen Forschung im internationalen Vergleich: Strukturen, Trends und Modelle, Wien , IHS Projektbericht, Juni 1998

Felt, U. (1999), Qualitätssicherung in Forschung und Lehre, in: Universität Wien/Logistisches Zentrum, Heide Pfennigbauer (Hg.), Studienpläne 2002, Positionen und Perspektiven der Reformdiskussion, Wien, S. 121-128 (4)

Galison, P. und B. Hevly (Hg.) (1991), Big Science. The Growth of Large-Scale Research, Stanford

Gibbons, M., C. Limoges, H. Nowotny, S. Schwartzman, P. Scott und M. Trow (1994), The New Production of Knowledge. The Dynamics of Science and Research, London

Gieryn, T. F. (1995), Boundaries of Science, in: S. Jasanoff, Gerald E. Markle, James C. Petersen, Trevor Pinch (Hg.), Handbook of Science and Technology Studies, Thousand Oaks/London/New Delhi, S. 393-443

Goldsmith, M. (1986), The Science Critic. A Critical Analysis of the Popular Presentation of Science, London

Hornbostel, S. (1997), Wissenschaftsindikatoren. Bewertung in der Wissenschaft, Opladen, insbes. S. 237-320

House, Ernest R. (1993), Professional Evaluation – Social Impact and Political Consequences, London

Merton, R. K. ([1973] 1985a), Die normative Struktur der Wissenschaft, in: ders., Entwicklung und Wandlung von Forschungsinteressen. Aufsätze zur Wissenschaftssoziologie, Frankfurt/Main, S. 86-99

Merton, R. K. ([1973] 1985b), Der Matthäus-Effekt in der Wissenschaft, in: ders., Entwicklung und Wandlung von Forschungsinteressen. Aufsätze zur Wissenschaftssoziologie, Frankfurt/Main, S. 147-171

Price, D. de Solla ([1963] 1974), Little Science, Big Science. Von der Studierstube zur Großforschung, Frankfurt/Main

Rossiter, M. W. (1993), The ~~Matthew~~ Matilda Effect in Science, in: Social Studies of Science, 23, S. 325-341

Travis, G. D. L. und Collins, H. (1991), New light on old boys: cognitive and institutional particularism in the peer review system, in: Science, Technology and Human Values, 16 (3), S. 322-341

Trow, M. (1994), Academic reviews and the culture of excellence. Studies of Higher Education and Research 2, Stockholm, The Council for Studies of Higher Education

Van Vught, F. A. (1997), The Humboldtian University under pressure – New forms of quality review in Western European higher education, in: Altrichter, H. et al. (Hg.), Hochschulen auf dem Prüfstand, Innsbruck, S. 48-87

Weinberg, A. M. (1963), Criteria for Scientific Choice, in: Minerva 1 (2), S. 159-171

Weinberg, A. M. (1989), Criteria for evaluation, a Generation Later, in: Ciba Foundation Conference (Hg.), The Evaluation of Scientific Research, Chichester, S. 3 -15

Weingart, P. und M. Winterhager (1984), Die Vermessung der Forschung. Theorie und Praxis der Wissenschaftsindikatoren, Frankfurt/New York

Weingart, P., R. Sehringer und M. Winterhager, Which reality do we measure? in: Scientometrics 19 (5-6), S. 481-493

Weingart, P., R. Sehringer und M. Winterhager (Hg.) (1991), Indikatoren der Wissenschaft und Technik: Theorie, Methoden, Anwendungen, Frankfurt/Main

Wenneras, C. und Wold, A. (1997), Nepotism and sexism in peer-review, in: Nature 387, S. 341-343

Woolgar, S. (1991), Beyond the citation debate: Towards a sociology of measurement technologies and their use in science policy, in: Science and Public Policy 18 (5), S. 319-326

Ekkehard Nuissl von Rein

Unterschiedliche Aufgaben – gemeinsame Ziele?

Entwicklung und Bewertung der Leibniz-Institute

Die sogenannte Blaue Liste ist eine Liste derjenigen wissenschaftlichen Institute, die analog Artikel 91b des Grundgesetzes von Bund und Ländern gemeinsam gefördert werden. Den Namen hat die Liste von der Farbe des Papiers, auf dem die erste Gruppe von Instituten vor gut zwanzig Jahren aufgelistet wurde. Es handelte sich um 46 Einrichtungen in der Bundesrepublik Deutschland, die übrigens auch damals schon vom Wissenschaftsrat begutachtet wurden. Für die Blaue Liste sind damit übrigens bereits zu einer Zeit Evaluationsverfahren angewandt worden, als eine regelmäßige externe Bewertung von Forschungsleistungen noch kaum wissenschaftspolitische Aufmerksamkeit hatte.

Mit dem deutschen Einigungsprozeß Anfang der 90er Jahre hat sich – wiederum – der Wissenschaftsrat dafür eingesetzt, das als erhaltenswürdig eingestufte Forschungspotential der ehemaligen DDR in das gemeinsame Wissenschafts- und Forschungssystem zu überführen. Im Ergebnis wurden vierunddreißig Einrichtungen aus dem Osten Deutschlands in die Blaue Liste aufgenommen. Das Förderungsvolumen von Bund und Ländern – in der Regel jeweils zu 50% – für Einrichtungen der Blauen Liste beläuft sich seit dieser Zeit auf jährlich etwa 1,2 Milliarden DM. Die Blaue Liste war damit neben der Max-Planck-Gesellschaft, den Großforschungszentren der Helmholtz-Gemeinschaft und der Fraunhofer-Gesellschaft zum großen Verbund außeruniversitärer Forschung in Deutschland geworden.

Wichtig ist es, festzuhalten, daß das verbindende Element der Einrichtungen der Blauen Liste zu diesem Zeitpunkt (1993/1994) die Art ihrer Förderung war, die Bund-Länder-Finanzierung. Die Institute der Blauen Liste verband weder eine gemeinsame Grundidee noch ein historischer Gründungsakt noch ein Kanon übergreifender Regeln noch eine gemeinsame Forschungskonzeption. Mögliche andere Aspekte von Gemeinsamkeit waren damals nicht diskutierbar, da die Institute untereinander nur minimal kommunizierten, geschweige denn kooperierten, und vielfach unbekannt war, welche Institute zum Förderungsinstrument der Blauen Liste gehörten. Gemeinsame Probleme vor allem im Bereich der Finanzierung wurden in einer Arbeitsgruppe der Blaue Liste-Institute besprochen, die ohne materielle Basis einen lockeren Zusammenschluß interessierter Institutsvertreter darstellte.

Am 11. April 1994 hat die Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung den Wissenschaftsrat gebeten, bis Ende 1999 sämtliche Einrichtungen der Blauen Liste zu evaluieren. Der Wissenschaftsrat kam dieser Bitte nach, richtete – mit Mitteln aus Instituten der Blauen Liste – eine Arbeitsgruppe zur Evaluation und einen eigenen Ausschuß ein, erarbeitete einen Fragenkatalog und ein Bewertungsverfahren und begann Anfang 1995 mit der Evaluation der Institute. Die

vom Wissenschaftsrat erarbeiteten Stellungnahmen sind mittlerweile in drei Bänden veröffentlicht, mindestens ein vierter wird noch erscheinen.

Der Beschluß, die Institute der Blauen Liste zu evaluieren, hatte unverzüglich Konsequenzen. Der Zusammenschluß zu einer eigenen Wissenschaftsgemeinschaft, die zuerst Wissenschaftsgemeinschaft Blaue Liste hieß, bis sie sich in Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz (WGL) umbenannte, ist unmittelbar auf den Evaluationsbeschluß und das beginnende Evaluationsverfahren zurückzuführen. Die Gründe für den Zusammenschluß, die Ziele der Wissenschaftsgemeinschaft, die Struktur in Sektionen und die Diskussionen in den vergangenen Jahren belegen den engen Zusammenhang dieser forschungspolitischen Aktivierung mit der Evaluation. Die Evaluierung der Institute hatte daher, bevor sie überhaupt begann, eine erste unmittelbare Wirkung, die man vermutlich unter „Impact" einordnen muß. Die Besprechungen und Versammlungen der Vertreter der Blaue Liste-Institute waren geprägt von Unsicherheiten, was die anstehende Evaluation bedeutet, von Zweifeln über den Sinn eines Zusammenschlusses bis hin zum energischen Verfechten der forschungspolitischen Notwendigkeit zur Gründung einer Wissenschaftsgemeinschaft. Der Zuschnitt der Sektionen folgte bereits den ersten erkennbaren Evaluationskriterien, und auch die Formulierung der Aufgaben der Wissenschaftsgemeinschaft ist von diesen beeinflußt.

Die ersten Gespräche in den Sektionen, im Plenum und in den informellen Treffen dazwischen drehten sich immer wieder um das Evaluationsverfahren, das Frageraster und – vor allem – erste Erfahrungen mit Institutsbegehungen. Insbesondere in den Sektionen hingen nervöse Institutsleiter (es handelt sich ausschließlich um Männer) gebannt an den Lippen der erleichterten Kollegen, die über erfolgreich abgeschlossene Begehungen berichteten. Dabei war im Laufe der Zeit durchaus ein Professionalitätsschub erkennbar: Die Berichte wurden immer präziser und strukturierter, die Nachfragen immer genauer und die Vorbereitungen für die Evaluation und die Begehung immer zielgerichteter. Was auch immer man zu den Wirkungen des Evaluationsverfahrens bezüglich der Qualität der Institute sagen mag: Die Qualität der professionellen Bewältigung von Evaluationsverfahren hat sich in den Instituten der ehemaligen Blauen Liste mittlerweile auf ein exzellentes Niveau bewegt.

Der Zusammenschluß in der heutigen Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz hat allerdings, dies sei bereits jetzt betont, mittlerweile eine eigene Dynamik entfaltet. Sie wird unterstützt durch eine zwar kleine, aber effektiv arbeitende Geschäftsstelle, eine sinnvoll strukturierte Gremienarbeit und eine Vielzahl neugeknüpfter Kontakte und Kooperationen zwischen den Instituten. Ich kann dies beispielsweise für den Bereich der Bildungsforschung sagen: Hier haben die Leibniz-Institute bereits auf verschiedenen Ebenen ein Netz geknüpft (etwa bei Umweltbildung, neuen Medien in der Bildung, Aufbau von Kosten- und Leistungsrechnung) und sich inhaltlich-konzeptionell auf eine gemeinsame Konferenz im Jahre 2001 zu Fragen einer künftigen Bildungsgesellschaft verständigt. Diese Zusammenarbeit ins-

besondere vom Deutschen Institut für Erwachsenenbildung, Deutschen Institut für Internationale Pädagogische Forschung, Institut für Pädagogik und Naturwissenschaften und Institut der Deutschen Sprache ergab sich gewissermaßen naturwüchsig aus der engen Zusammenarbeit innerhalb der Sektion A der WGL: Bildungsforschung und Museen. Die anfängliche Unverbundenheit der Leibniz-Institute hat sich bereits jetzt zu einem kooperativen Netzwerk interdisziplinärer Institute mit eigenem Profil und Selbstverständnis entwickelt. Das „Memorandum“ vom März 1999 (WGL 1999) legt davon Zeugnis ab.

Zurück zur Evaluation: Rufen wir uns noch einmal die Ziele des Evaluierungsverfahrens in Erinnerung. Sie bestehen darin, „die Wissenschaft insgesamt zu fördern und Flexibilität und Innovationskraft der in der Blauen Liste geförderten Institutsgemeinschaft zu stärken“ (Dagmar Schipanski, Vorsitzende des Wissenschaftsrates, in: Wissenschaftsrat 1996, S. 6). Es wurden mittlerweile von unterschiedlichen Beteiligten noch andere Ziele genannt, wie etwa dasjenige, daß eine Hilfe für die Zielorientierung der Institute gegeben werden, Transparenz geschaffen und die gute Qualität der Arbeit belegt werden solle.

Es ist nicht daran zu zweifeln, daß diese Ziele verfolgt werden und bei dem Beschluß, ein solch aufwendiges Evaluierungsverfahren von über achtzig Instituten vorzunehmen, eine Rolle gespielt haben. Sie sind aber zu ergänzen und in einen größeren forschungspolitischen Kontext zu stellen. Dieser betrifft die Höhe der aufgewendeten Mittel für die Institute der Leibniz-Gemeinschaft, die Frage der gemeinsamen Förderung von Bund und Ländern und damit weitergehend das föderale Strukturprinzip, die Zuordnung der Institute zur Leibniz-Gemeinschaft anstatt zu einer Hochschule oder zu einer anderen Wissenschaftsorganisation (wie insbesondere Max-Planck- und Fraunhofer-Gesellschaft), die öffentliche Legitimation des verausgabten Geldes und schließlich das Prinzip der Flexibilität – nicht als ein Prinzip der Arbeit der einzelnen Institute, sondern als ein Prinzip des staatlichen Förderungsinstruments.

Flexibilität bedeutet ja forschungspolitisch nicht nur eine Verschiebbarkeit eines Instituts zwischen übergeordneten Forschungsorganisationen, sondern vor allem auch das Ausscheiden aus und die Neuaufnahme von Instituten in Förderungsinstrumentarien. Entsprechend wird bei der Blauen Liste heute vielfach von einem „Omnibus“-Prinzip gesprochen, das die angestrebte Flexibilität veranschaulicht: Ein Omnibus fährt einen vorgeschriebenen Weg, auf dem Personen zu- und aussteigen. Evaluationsverfahren werden in diesem Kontext zu einem Instrument der Lenkung von Forschung, die sich auf die Überzeugungskraft der wissenschaftlichen Ergebnisse kapriziert. Die Evaluierung der Blaue Liste-Institute gehört damit – in der Klassifizierung von Guba und Lincoln (1989) – zur vierten Generation von Evaluierungen, derjenigen mit Wirkungs-Orientierung, nach den vorangegangenen Generationen der reinen Kosten-Nutzen-Analyse, der Prozeß-Orientierung und der Methodenorientierung. In dieser vierten Generation spielen die Betroffenen und ihre Partizipation am Evaluierungsverfahren eine große Rolle, aber auch der explizite

Verzicht auf eindeutige wissenschaftliche Objektivität: „Führt man sich die Entwicklungen der Evaluationspraxis in den letzten Jahren vor Augen, so läßt sich feststellen, daß sich diese zwischen einer wissenschaftlichen grundlagenorientierten Forschung und einer mehr praxisorientierten Organisations- und Institutionsberatung ansiedeln läßt: In Abgrenzung zur wissenschaftlichen Forschung ist Evaluation weniger auf neutrale Ergebnisermittlung durch die exakten Anwendungen empirischer Forschungsmethoden ausgerichtet, sondern ist ein handlungsorientiertes, auf Veränderungen ausgerichtetes Verfahren, das ausdrücklich Bewertungen einschließt, die in einem herkömmlichen mit wissenschaftlicher Objektivität verbundenen Sinne nicht zulässig sind“ (Liebald 1996, S. 249).

Es geht also nicht um eine wissenschaftliche Bewertung, sondern um ein diskursives Verfahren unter weitestgehendem Einbezug von Kompetenz und mit zunehmender Präzision und Akzeptanz. Eine Kritik des vorgenommenen Evaluierungsverfahrens vom Standpunkt wissenschaftlichen Arbeitens würde daher von Grundlage und selbstdefinierten Verfahren fehlgehen. Es handelt sich um eine forschungspolitische Bewertung.

Entsprechend ist auch das Evaluierungsverfahren der Leibniz-Institute aufgebaut. Vorgegangen wird in drei unterschiedlichen Stufen:

- In der ersten Stufe wird die wissenschaftliche Qualität des Instituts und seiner Arbeit durch eine Expertengruppe bewertet, die vom Wissenschaftsrat zusammengestellt wird und das Institut nach vorhergehender ausführlicher schriftlicher Information „begeht“. Diese Expertengruppe ist Autor der „Peer-group-review“ und verfaßt weitgehend im Konsens einen Bewertungsbericht, der anschließend nicht mehr verändert werden kann.
- Die zweite Stufe erfolgt im Wissenschaftsrat, zunächst im Ausschuß Blaue Liste, sodann im Wissenschaftsrat selbst. Dort wird nach dem Grundsatz, die wissenschaftliche Qualität des Instituts sei eine notwendige, aber keine hinreichende Bedingung für das Mitfahren im Omnibus, diskutiert über die Bewertung des Berichts der Expertengruppe hinsichtlich wissenschaftspolitischer, gesamtstaatlicher und überregionaler Aspekte. Wichtig ist dabei etwa die Frage, ob und warum die Arbeit nicht an einer Hochschule geleistet werden kann. Wer die vorliegenden Evaluationsberichte aufmerksam liest, kann feststellen, daß hier die Expertenberichte in einen teilweise anderen Kontext gestellt und teilweise andere Urteile gefällt werden.
- Die dritte Stufe schließlich liegt in der Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung, die auf der Grundlage der Empfehlung des Wissenschaftsrates die Entscheidung über Weiterförderung oder Ausscheiden eines Instituts trifft. Hier spielen Interessen eine Rolle, die man als Kriterien in den zuvorgehenden Stufen des Evaluierungsverfahrens vergeblich sucht, etwa länderspezifische strukturpolitische Interessen, Fragen des Finanzausgleichs zwischen Bund und Ländern und Fragen der Profilierung von Regionen und

Politiken. Der Weg von der Bund-Länder-Kommission über den Wissenschaftsrat in die evaluierende Expertengruppe geht also – mit entsprechenden Unwägbarkeiten – wieder zurück in die Bund-Länder-Kommission.

Das Evaluierungsverfahren von Bund-Länder-Kommission und Wissenschaftsrat bei den Instituten der Blauen Liste ist groß angelegt und hat einen hohen Anspruch. Die Systematik, mit der es konzipiert, operationalisiert und seit einigen Jahren durchgeführt wurde, sucht ihresgleichen. Wir sind heute in einer Situation, in welcher der Großteil der Leibniz-Institute nach diesem Verfahren evaluiert worden, „durch das Fegefeuer der Evaluation, der externen Bewertung durch den Wissenschaftsrat" (Lange 1999) gegangen ist. Sowohl die Verfahren bei der Erhebung von Daten zu den einzelnen Instituten als auch die Verfahren der Bewertung und der darauf aufbauenden Entscheidung stellen einen Markstein in allen in Deutschland bekannten Verfahren zur Evaluation dar. Es ist bezeichnend, daß die zugrunde liegenden Kriterien ebenso wie das Evaluationsverfahren im Grundsatz unstrittig sind – dies ist gerade angesichts der im Einzelfall weitreichenden negativen Konsequenzen eine bemerkenswerte Tatsache. Es ist davon auszugehen, daß am Ende des Evaluationsverfahrens im Jahre 2000 „eine flexible, sehr moderne Wissenschaftsorganisation vor uns steht", die „dann als einzige Forschungsorganisation nach strengen Kriterien durchleuchtet worden" ist (Lange 1999).

Sieht man und beurteilt man dieses Ergebnis, so ist es um so höher zu bewerten, als es auf der Grundlage von Kriterien- und Beurteilungssystemen zustande kommt, die immer im Konkreten mit schwierigen Fragen der Meßbarkeit, der Festlegung von Merkmalen und der Akzeptanz der Beurteilungen zu kämpfen hat. Dies betrifft sowohl das Objekt der Evaluation, das evaluierte Institut, wie auch das Subjekt, die begutachtende „Peer-Group" und die sich an diese anschließenden Instanzen des Wissenschaftsrates und der Bund-Länder-Kommission.

In der Evaluation des Objekts, also der Institute, sind es im wesentlichen vier Aspekte, die einer besonderen Diskussion bedürfen:

- die Merkmale des Instituts bzw. die sie erfassenden Kriterien;
- die Produkte des Instituts;
- die Meßgrößen für Leistung und Qualität sowie
- die Analyse der Organisation.

Zu den Kriterien und Merkmalen:

Unterschiedliche Aufgaben – gemeinsame Ziele, – so heißt mein Beitrag, mit einem Fragezeichen versehen. Im Verlaufe des Evaluierungsverfahrens wurde ein Fragebogen eingesetzt, mit dessen Hilfe nicht nur die Unterlagen und Kennzahlen der Institutionen erhoben wurden, sondern an denen entlang auch die Bewertung erfolgte. Zunächst fällt bei diesem Fragebogen auf, daß er nur einen minimalen

Unterschied zwischen Forschungs- und Serviceinstituten macht. Dies ist, wenn man Ziel und Struktur dieser unterschiedlichen Institutstypen nimmt, nicht unproblematisch. Bei Serviceinstituten etwa ist die Frage der Nutzung, der Adressaten, der Dissemination, der Vernetzung und der Kooperation (über die Produkte spreche ich gleich) anders einzuschätzen als bei Forschungsinstituten. Das Verschlagworten wissenschaftlicher Aufsätze etwa ist eine gänzlich andere Arbeit als ihr Verfassen, was in den Fragebögen nicht angemessen erfaßt wird.

Dies gilt aber auch für die Unterschiede, welche jeweils im natur-, sozial- und geisteswissenschaftlichen Bereich Forschung charakterisieren. So sind etwa die Finanzmittel im naturwissenschaftlichen Bereich anders zu gewichten als im geisteswissenschaftlichen Bereich, Veröffentlichungen in bezug auf nationale oder internationale Akzeptanz gänzlich anders einzuschätzen beim Institut für Deutsche Sprache in Mannheim als beim Neurobiologischen Forschungsinstitut in Magdeburg. Die Fragebogen lassen, in der bisherigen Struktur, kaum eine institutsspezifische Modifikation zu. Entsprechend lesen sich auch die Darstellungen, welche die Institute von sich selbst analog zu den Kriterien des Fragebogens geben. Ein Erhebungsinstrument mit Kernfragen und einem Satz modifizierbarer Kranzfragen wäre der unterschiedlichen Situation und Gestalt der Institute sicher angemessener und im Evaluierungsverfahren auch leichter einsetzbar.

Der Fragebogen bzw. die Liste der erfaßten Merkmale ist somit derzeit umfassend und klar definiert. Das Grundsatzproblem, daß Kriterien zur angemessenen Bewertung erfüllter Aufgaben aufgabenabhängig deduziert werden müßten, dies jedoch mit einem fachübergreifenden Beurteilungssystem konfligiert, ergibt jedoch immer die Notwendigkeit und Möglichkeit, Kriterien und Merkmale auf eine noch bessere Planung zu den Realitäten der Institute hin zu überprüfen.

Zum Produktbegriff:

Die Definition wissenschaftlicher Produkte ist – und dies gilt für alle Fachdisziplinen – objektiv nicht leistbar. Sie ist ähnlich schwierig wie etwa die Definition des Produkts von „Bildung“. Produktdefinitionen in Evaluierungsverfahren wie dem hier vorliegenden sind immer nur Näherungswerte, in denen versucht wird, möglichst eng an die Realität der geleisteten Arbeit heranzukommen. Interessanterweise – und dazu schlagen in den letzten zwei bis drei Jahren die Diskussionen teilweise hoch – nähert sich die evaluierungsorientierte Produktdefinition derjenigen bei der Einführung einer Kosten- und Leistungsrechnung. Auch dort werden, um Kostenstellen und Kostenarten den Kostenträgern zuordnen zu können, definierte Produkte benötigt. Die Diskussion dort entzündet sich vor allem am Begriff der *Leistungs*rechnung, weil vielfach unter Leistung in diesem Kontext die Bewertung der Qualität verstanden wird (was gar nicht intendiert ist). Alleine über die Dimension der Kosten ist Leistung nicht bewertbar.

Eine andere Diskussion verdeutlicht ebenfalls die Problematik, das wissenschaftliche Produkt zu definieren, so die Frage, ob Politikberatung streng genommen Produkt eines Forschungsinstituts sein kann (insbesondere die wirtschaftswissenschaftlichen Institute, allesamt auf der Blauen Liste und seit vielen Jahren als „Wirtschaftsweise" politikberatend tätig, gerieten in diese Diskussion), oder die Frage, inwieweit Forschung in Museen von Ausstellungen abzugrenzen ist. Zu beiden Aspekten hat der Wissenschaftsrat dankenswerterweise zusammenfassende Stellungnahmen abgegeben, die mehr Klarheit in die Diskussion bringen.

Vielfach wurde die Produktfrage so gelöst, daß alles das, was von Seiten des Instituts extern abgegeben wird, als Produkt definiert wird, insbesondere natürlich Publikationen und Vorträge und Patente und Spezifika wie etwa Curricula, bestimmte Verfahren oder mediale Konserven. Mit einigem Recht könnte man die erfragten Aspekte wie Nachwuchsförderung, Lehrtätigkeit und Präsenz in Gutachter- und Fachgremien auch als Produkte verstehen.

Letztlich konzentriert sich jedoch der Produktbegriff auf die klassische Form von Vorträgen und Publikationen, auf deren *Bewertung* ich gleich noch eingehe. Die Heterogenität der Institute in Art und Gestalt ihrer Arbeit ist damit natürlich schwerlich abbildbar. Auch geben die so definierten Produkte weder Ansatzpunkte für Differenzierungen noch für Qualitätsmaßstäbe. Die Frage nach einem wissenschaftlichen Produkt als eine inhaltliche oder fachliche Kategorie darf dabei gar nicht gestellt werden, einmal aus den genannten systematischen Gründen, zum anderen natürlich deshalb, weil fachübergreifend nur formale Produktbegriffe verwendet werden können.

Zu Leistung und Qualität:

Wissenschaftliche Leistung und Qualität sind nun diejenigen Kategorien, die bei der Bewertung die zentrale Rolle spielen. Dazu werden im wesentlichen drei Meßsysteme verwendet: zum ersten das fachliche Expertenwissen der Peer-group, zum zweiten formale und gewissermaßen objektivierte Kennzahlensysteme und zum dritten Meßkategorien der Akzeptanz im Felde, insbesondere die eingeworbenen Drittmittel.

Das Expertenwissen der Peer-group ist zweifellos eine wichtige Ressource zur Qualitätsbeurteilung; die besonderen Probleme dabei spreche ich noch an, wenn ich zum Subjekt des Evaluierungsverfahren komme. Die Kennzahlen richten sich hauptsächlich auf das Verhältnis von Personal, aufgewandten Mitteln und erstellten Produkten und konzentrieren sich dabei auf letztere, die erbrachten Produkte. Ein besonders beliebtes Meßinstrument ist dabei die Bibliometrik, also die Menge von Publikationen in nationalen und internationalen referierten Fachzeitschriften. Hier wird, gerade im naturwissenschaftlichem Bereich, mit höchst differenzierten Quotienten gearbeitet, welche die wissenschaftliche Akzeptanz der Produkte in

diesen referierten Systemen belegen sollen. Vielfach wird argumentiert, die erforderliche Exzellenz fehle, weil der Quotient zu niedrig oder zu viele wissenschaftliche Beschäftigte an seinem Zustandekommen beteiligt seien.

Wir haben hier ein System der Objektivierung von Peer-group-Verfahren. Auch referierte Fachzeitschriften sind nichts anderes als Peer-group-Systeme, auch für sie gelten die blinden Flecken dieses Systems. Darüber hinaus besagt die Menge von Publikationen und Vorträgen nach wie vor nichts über deren Qualität. Es ist in der Wissenschaftssoziologie belegt, daß das Prinzip des „publish or perish" vielfach zur Zerstückelung von Arbeitsergebnissen und damit eher qualitätsmindernd als qualitätssteigernd wirkt. Auch erfaßt die Bibliometrik viele Produkte und Leistungen gar nicht, weil sie in anderen Formen entstehen (übrigens neuerdings auch immer häufiger im Internet, das bibliometrisch noch gar nicht zählt) oder von anderer Qualität sind (z. B. Verschlagwortung). Und schließlich hat jede Disziplin ihre eigene Publikationsstruktur, in Naturwissenschaften etwa vorwiegend Aufsätze in Zeitschriften, in Sozialwissenschaften Beiträge in Sammelbänden, in Rechtswissenschaften Kommentare und Monographien.

Die dritte Kennzahlen-Systematik für Leistung und Qualität ist die Einwerbung der Drittmittel. Hier gibt es eine „implizierte" Hierarchie: Am wenigsten gelten Drittmittel aus der staatlichen Ressortforschung, besser bewertet sind Drittmittel aus der Wirtschaft und dem Verkauf eigener Produkte, noch höher bewertet sind Drittmittel aus internationalen Quellen, und am höchsten bewertet sind Drittmittel der Deutschen Forschungsgemeinschaft (DFG). Einmal abgesehen davon, daß diese Einschätzungshierarchie keineswegs immer dem Verhältnis von Aufwand und Ertrag bei der Drittmittel-Akquisition entspricht, zeigt vor allem auch die Bewertung der DFG die Tücke des Systems. Zunächst handelt es sich bei der DFG, dies muß nicht weiter erläutert werden, wiederum um ein Peer-group-System mit den noch genauer zu nennenden Problemen. Zum zweiten geht die Förderung der DFG im großen und ganzen aus von Empfängern, welche eine hochschulähnliche Struktur aufweisen – also etwa keine eigenen Mietkosten haben und Menschen in hochschulähnlichen Beschäftigungsverhältnissen arbeiten lassen. Für Leibniz-Institute ist dies vielfach ein inhaltliches und ein Refinanzierungsproblem. Schließlich tendiert die Begutachtungssystematik der DFG dazu, dasjenige, was Leibniz-Institute als übergreifenden Nenner auszeichnet, geringer zu würdigen: einerseits Interdisziplinarität und andererseits Anwendungs- und Problemorientierung. Die Leibniz-Institute haben zwar von allen außeruniversitären Forschungsgemeinschaften den engsten Bezug zu Universitäten, konnten aber – zumindest in einzelnen Fachgebieten – diesen strukturellen Nachteil nicht ausgleichen.

Insgesamt stammt die Wertigkeitshierarchie der Drittmittel vor allem aus einer universitären Sichtweise, welche die spezifischen Arbeitsweisen- und Aufgabenorientierungen außeruniversitärer wissenschaftlicher Institute nur unzureichend in den Blick nimmt.

Zur Organisation:

Viele der Kriterien und Fragen bei der Begutachtung zielen ab auf die innere Organisationsstruktur der Institute. So wird etwa gefragt nach der Kooperation zwischen den Abteilungen, nach der Altersstruktur, nach dem Zusammenwirken zwischen Institutsleitung und wissenschaftlichem Beirat, nach Auswahl und Qualifikation der Beschäftigten usw. Vielfach wird mit Begriffen gearbeitet wie Matrixstruktur, Aufgabenkonzentration und effektiver Steuerung. Diesen Komplex sowohl des Fragenkatalogs als auch des Umgangs mit ihm in den Bewertungsberichten finde ich am problematischsten. So werden etwa Strukturen von Instituten als zu zergliedert oder zu groß bewertet, ohne daß erkennbar ist, welche Modelle und Konzepte einer Organisation dahinter stehen. Auch ist erkennbar, daß die Fragen von Kosten einerseits und wissenschaftlicher Leistung andererseits in keinen systematischen Bezug zur Organisationsstruktur gesetzt werden. Es mag sein, daß bei der Erstellung des Fragebogens der Blick zu sehr auf die wissenschaftlichen Ergebnisse, weniger auf die Bedingung ihres Erstellens gerichtet war. Es mag auch sein, daß hier aus dem Blick einer (traditionellen) Hochschule wenig Empathie für die betriebsförmige Organisation eines Instituts vorlag. Wie auch immer: Vom Ergebnis her scheinen mir gerade in diesem Bereich die Beurteilungen der Institute zwar in vielen Fällen erstaunlich eng an den betrieblichen Problemen, vielfach jedoch ohne eine organisationssoziologische oder betriebswirtschaftliche konzeptionelle Grundlage zu sein.

Nun zur anderen Seite des Evaluierungsverfahrens, dem evaluierenden Subjekt. Hier sind es vor allem drei Aspekte, auf die ich eingehen möchte:

- die Expertengruppe selbst,
- den von ihr erstellten Text sowie
- Prozeß und Dynamik des Evaluierens.

Zur Expertengruppe, der „Peer-group“:

Bereits bei ihrer Zusammensetzung durch den Wissenschaftsrat sind wichtige Entscheidungen zu treffen. Zum einen soll die Expertengruppe mehrheitlich fachnah zur Institutsarbeit, zum anderen aber nicht zu eng mit dem Institut verbunden sein. In den meisten Fällen lassen sich schwer Expertinnen und Experten finden, die fachlich einschlägig arbeiten, zum Institut aber nicht bereits in einer definierten Beziehung stehen. Vielfach sind sie in einem der Ausschüsse und Gremien des Instituts Mitglied, die exzellentesten Fachvertreter und Fachvertreterinnen in der Regel auch im wissenschaftlichen Beirat. Vielfach stehen sie auch in anderen Interessenkontexten zum Institut. Und schließlich haben gerade Fachkolleginnen und Fachkollegen immer eine bestimmte Auffassung von der Arbeit des Instituts, die sie gewissermaßen als Vorurteil in die Evaluierung mitbringen. Die Kooptation einiger

fachferner Mitglieder in der Peer-group ist dabei hilfreich und vielfach recht erfrischend, erhöht aber wiederum gewisse fachliche Unwägbarkeiten.

Die Definition einer Peer-group aus der eigenen „Scientific community" schließt Fehlurteile nicht aus. So argumentiert etwa Fischer (1998) mit historischen Belegen, daß das System von Peer-reviews immer schon zum Mittelmaß hin tendierte und exzellente wissenschaftliche Leistungen eher ausschloß. Die Beispiele Freud und Benjamin sind hier eher bekannt, aber auch das von ihm genannte Beispiel Galilei gibt insofern zur Nachdenklichkeit Anlaß, als Galilei erst im hohen Alter zu publizieren begann. Neben diesem generellen Problem der Tendenz zum kleinsten gemeinsamen Nenner (Mittelmaß) existiert natürlich das Problem der Konkurrenz etwa um Forschungsmittel oder auch grundsätzlich um staatliche Mittel, die gelegentlich etwa von Hochschulleuten als in zu hohem Ausmaß in außeruniversitäre Forschungseinrichtungen investiert gesehen werden. Schließlich bedeutet fachdisziplinärer Sachverstand tendenziell immer die Schwierigkeit, interdisziplinäre Ansätze ausreichend zu würdigen; Fragen und Urteile werden eher aus dem Blickwinkel der eigenen Disziplin gestellt und gefällt.

Um so bemerkenswerter ist es, festzustellen, daß die von einer systematischen Blickweise her bestehenden Probleme einer Peer-group in den bisherigen Evaluierungen von Leibniz-Instituten nur im Ausnahmefall erkennbar, aber in keinem mir bekannten Fall durchschlagend waren. Die große Sorgfalt, die auf die Zusammenstellung und die Arbeitsweise der Peer-groups gelegt wird, zahlt sich hier als eine ausgewogene Beurteilung und eine unstrittige Akzeptanz derselben aus.

Zum Text:

Der Bewertungstext, den die Peer-group erstellt und der in der Leibniz-Evaluierung als nicht mehr veränderbares Produkt in die weiteren Stufen des Verfahrens eingeht, dieser Text schließlich ist ein Kunstwerk. Er ist ein Kunstwerk insofern, als in ihm die aus unterschiedlichen Blickwinkeln genannten Aspekte zu einem systematischen Gesamttext kompiliert werden, der in der Gruppe abgestimmt und abgesegnet wird. Es ist verständlich, daß dieses Verfahren sich über Monate hinzieht. Die Texte haben, gemessen an ihrer Entstehung als Gruppenprodukt, durchweg eine erstaunliche Konsistenz.

Dennoch gibt es Schwierigkeiten, die zu Beginn des Evaluierungsverfahrens noch häufiger auftraten als nun, nachdem einige Erfahrungen gesammelt wurden. Ein Problem besteht zunächst darin, daß die Berichte die Institute aus der Sicht von Fachdisziplinen erfragen und bewerten. Auch im Gruppenprodukt kann diese Disziplinorientierung teilweise erkannt werden. Die Texte zeigen, daß eine interdisziplinäre Arbeit des Instituts von der disziplinorientierten Herangehensweise nur schwer angemessen erfaßt wird. So entsteht etwa eher die Nebeneinanderstellung fachdisziplinärer Monita als ein integrativer interdisziplinärer Zugang. Ein zweiter Aspekt der

Texte liegt im Stil: Erklärt werden oft nicht positive Eindrücke (sie werden allenfalls, wenn überhaupt, benannt), erklärt werden Probleme. Probleme gewinnen daher in den Evaluierungstexten in der Regel einen wesentlich höheren Stellenwert als positive Eindrücke. In den Diskussionen des Ausschusses Blaue Liste des Wissenschaftsrates wird eben dieser Sachverhalt seit Beginn des Evaluierungsverfahrens immer wieder thematisiert, weil diese Textstruktur vielfach gar nicht der Bewertungsintention entspricht.

Schließlich zum Prozessualen:

Das Vorgehen der Evaluierungsgruppe ist ein rückwärtsgewandtes Vorgehen. Ermittelt werden Fakten und Daten aus den vergangenen drei bis fünf Jahren, diskutiert und begutachtet wird der Status quo. In diesem Verfahren steckt eine Dynamik insofern, als die Institute sich auf die Evaluierung einstellen und mit zunehmender Präzision die Kriterien erfüllen. Aber auch im Vorfeld der Evaluierung erfolgen vielfache Kommunikation, Kontakte und Diskussionen. Institute befinden sich nicht im luftleeren Raum. Wesentlich aber, und darum spreche ich diesen Punkt an, ist der rückwärts gerichtete Blick. Die Evaluierung zielt letztlich in die Zukunft, Zielorientierungshilfe und wissenschafts- und forschungspolitische Einschätzung verfolgen Ziele für die Zukunft, weniger die Bewertung der Vergangenheit. Der Prognosewert des bewerteten Status quo ist – zumindest nicht immer – übermäßig hoch. Fischer spricht hier von einer „Zukunftsblindheit des Bewertungssystems“ – so kraß würde ich das nicht formulieren, es besteht aber ein zu definierendes Verhältnis von Zukunftsorientierung und Evaluation des Status quo.

Mit der Zukunftsorientierung sind wir auch bei dem Stichwort, welches den forschungspolitischen und organisatorischen Kontext des Evaluationsverfahrens der Blauen Liste aufruft. Man kann als Vertreter der Leibniz-Gemeinschaft sehr zufrieden sein mit dem bisherigen Ergebnis der Evaluierung. Nur ganz wenigen Instituten wurde eine nur durchschnittliche oder gar schlechte wissenschaftliche Qualität bescheinigt, in nur ganz wenigen Fällen wurde beschlossen, die Institute im Rahmen der Blauen Liste nicht mehr weiter zu fördern. Nur in zwei Fällen erfolgte dies mit dem Hinweis, im Prinzip sei die Aufgabe forschungspolitisch nötig, nur das Institut erfülle sie nicht richtig. In der Regel wurde der negative Beschluß mit Verweis auf die Qualität der Arbeit gefällt. Dabei wurden kritische, aber durchaus nicht grundsätzlich abwertende Berichte der Bewertungsgruppe auf dem Weg über den Wissenschaftsrat bis hin zum Beschluß der Bund-Länder-Kommission teilweise einseitig ausgedeutet. Teilweise allerdings wurde ein neues Zukunftskonzept eingefordert, auf dessen Grundlage dann eine – reduzierte – Weiterförderung empfohlen wurde. Die einschlägigen Schließungsentscheidungen treffen etwa das Institut für Erdölchemie in Claustal/Zellerfeld, das Deutsche Bibliotheksinstitut in Berlin und das Deutsche Institut für Fernstudienforschung in Tübingen. Im Kontext der Reduktions-Entscheidungen liegen etwa das Deutsche Institut für Internationale

Pädagogische Forschung in Frankfurt/Main, das Hamburger Weltwirtschaftsarchiv und das Institut für wissenschaftlichen Film im Göttingen.

Wenig zufrieden kann ich als Vertreter der geistes- und sozialwissenschaftlichen Forschung sein, was die inhaltlichen Konsequenzen der negativen Beschlüsse betrifft. Nicht unbedingt in absoluten Zahlen, aber doch in Relation zu den anstehenden Entscheidungen sind vor allem Institute der geistes- und sozialwissenschaftlichen Richtung betroffen. So ist etwa im Bildungsbereich mit der Halbierung des Instituts für den Wissenschaftlichen Film, der Reduktion und Funktionsänderung des Deutschen Instituts für Internationale Pädagogische Forschung in Frankfurt/Main und der Schließung des Deutschen Instituts für Fernstudienforschung in Tübingen fast die Hälfte der einschlägigen Kapazitäten gekappt worden. Neu aufgenommen dagegen werden in die Blaue Liste zwei naturwissenschaftliche Institute. Hier handelt es sich demnach – implizit und teilweise auch explizit – um Richtungsakzente der Forschungslandschaft, welche über die Bewertung einzelner Institute hinausgehen.

Damit sind wir bei dem Kontext von Evaluation, der bislang hinter der einzelnen Institutsbewertung zurücktritt: der systemischen Evaluation. Die bislang vorliegenden Ansätze einer systemischen Evaluation im Rahmen des laufenden Evaluierungsverfahrens sind die Stellungnahme des Wissenschaftsrates zu den Physikinstituten der Blauen-Liste in Berlin unter dem Stichwort "Übergreifende Gesichtspunkte“, zu den Wirtschaftsforschungsinstituten der Blauen Liste in den alten Ländern und zu den Museen. Bei den Wirtschaftsforschungsinstituten ging es um ihr Verhältnis zueinander sowie vor allem um die Frage der Bewertung des Produkts „Politikberatung“ als eines wissenschaftlichen Produkts. Bei den Physikinstituten ging es vor allem um das Verhältnis von Grundlagenforschung und Anwendungsbezug sowie zu den Kooperationen und zur gemeinsamen Verwaltung im Forschungsverbund Berlin. Bei den Museen ging es um die Frage der Bewertung und Abgrenzbarkeit von Forschung. Deutlich wird in den drei Fällen, daß der kommende Schritt die systemische Evaluation wird sein müssen, welche über die einzelnen Institute hinausgeht.

Die Institute der Leibniz-Gemeinschaft sind, auch dies eine Gemeinsamkeit, in ihrem Entstehen immer mit dem Versuch verbunden, ein bestimmtes Problem systematisch zu bearbeiten; sie enthalten daher in der Regel sowohl von der Genesis als auch von der Aufgabenstellung her einen Problem- und Anwendungsbezug. Dies entspricht natürlich keiner systematisch abgeleiteten Aufgabenstruktur, schon gar nicht innerhalb des Förderungsinstruments der Blauen Liste oder der Leibniz-Gemeinschaft. Es ist jedoch ein übergreifendes Charakteristikum. Eine systemische Evaluation der Leibniz-Institute wäre im wesentlichen daraufhin zu orientieren, eine Grundstruktur von Aufgabendefinition und Arbeitsweise, welche für die Institute gilt, in ein übergreifendes und systemisches Bewertungsraster umzusetzen. Dazu bedarf es jedoch eines zugrundeliegenden Konzeptes sowohl gesellschaftlicher Entwicklungsprozesse als auch forschungspolitischer Notwendigkeiten. Ein solches Konzept liegt bislang

nicht vor. Dies gilt jedoch nicht nur für die Leibniz-Gemeinschaft, sondern in gewisser Weise auch für die anderen drei, schon länger etablierten Forschungsorganisationen, für die – mit Ausnahme der Helmholtz-Gemeinschaft – seit kurzem ein Bericht zur Systemevaluation vorliegt. Vor allem aber, und dies ist eine forschungspolitisch für die gesamte Bundesrepublik erforderliche Aufgabe, fehlt es an einem Konzept von Forschungspolitik insgesamt, ein Konzept des systemischen Zusammenwirkens der großen Forschungsorganisationen.

Nun kann man entstehende und zu erforschende Probleme nicht durchweg prognostizieren oder gar in forschungspolitischen Entscheidungen vorwegnehmen. Forschungspolitische Konzeptionen bedürfen daher einer prozessualen Komponente, die Analysen und Erkenntnisse impliziert und Modifikationen zuläßt. Eine solche forschungspolitische Konzeption auf dem heutigen Stand wäre wohl Aufgabe des Wissenschaftsrates. Dieser – und insbesondere sein Ausschuß Blaue Liste – ist bislang mit dem laufenden Evaluierungsverfahren einzelner Institute mehr als ausgelastet. Er hat daher nur folgerichtig den Beschluß gefaßt, das Evaluierungsverfahren der Leibniz-Institute zukünftig der Leibniz-Gemeinschaft zu übertragen und sich seinerseits wieder auf konzeptionelle und übergreifende Aufgaben zu konzentrieren. Dies ist zweifellos aus übergeordneter forschungspolitischer Sicht zu begrüßen. Die Leibniz-Gemeinschaft selbst hat im Vorgriff auf diese kommende Aufgabe auch bereits einen Senat gebildet, der zukünftig solche Evaluierungsaufgaben erfüllen soll. Dieser Senat wird weitgehend analog zum bewährten System des Wissenschaftsrates arbeiten. Vor allem setzt er fort, was der Wissenschaftsrat und sein Ausschuß Blaue Liste bisher auch waren: Er ist eine Peer-group auf einer höheren Ebene.

Die bisherigen Beratungen zur Gestaltung einer künftigen Regelevaluation der Leibniz-Institute, vom Senat der WGL verantwortet, bestätigen die Eingangsfeststellung: Im Prinzip soll an dem begründeten und bewährten Verfahren des Wissenschaftsrates festgehalten werden. Diskutiert werden hauptsächlich – neben einer Überarbeitung des Fragebogens – zwei Punkte: die Zeiträume zwischen den (Regel-)Evaluationen (im Gespräch sind hier zwischen fünf und acht Jahre) und das Verhältnis der externen Evaluation zur laufenden begleitenden Evaluation der wissenschaftlichen Beiräte und Kuratorien (Modelle engerer oder weiterer Beteiligung an der externen Evaluation). Diese Beratungen sind sachlich und belegen das hohe Niveau, auf dem die Evaluationsdiskussion mittlerweile forschungspolitisch geführt wird.

Vermutlich wird erst auf der Basis eines übergreifenden Konzepts nicht nur für die Blaue Liste, in dem Ziel und Fahrstrecke des Omnibus Blaue Liste festgelegt werden, sondern für die Forschungsorganisationen in Deutschland insgesamt die derzeit noch implizite Einschätzung und Hierarchie der Forschungsorganisationen relativiert und diskutierfähig. Die Hierarchie der Forschungsorganisationen, in welcher die Max-Planck-Gesellschaft an der meist unausgesprochenen Spitze und die Leibniz-Gemeinschaft am unteren Rande steht, kann bislang im wesentlichen als Vorurteil erklärt werden, welches historisch begründbar ist, aber den konkreten Realitäten der

Arbeit nicht entspricht. Es ist aber auch dadurch erklärbar, daß die Ziele und Aufgaben der Forschungsorganisationen im Verhältnis zueinander nur dann vernünftig einzuschätzen sind, wenn sie im Rahmen einer übergreifenden Forschungskonzeption systemisch zuordenbar sind.

Die Leibniz-Gemeinschaft hat hier in der Zukunft eine große Aufgabe. Sie hat fortzusetzen, was sie bereit begonnen hat: die kooperative Vernetzung der Institute, die Profilierung des typischen Leibniz-Instituts, das Profil einer Forschungsgemeinschaft, die über Anwendungsbezug, Interdisziplinarität und Hochschulkooperation einen unverzichtbaren Anteil an der deutschen Forschungslandschaft hat. Dazu wird die Leibniz-Gemeinschaft ihre zentralen Funktionsteile verstärken (z. B. über einen Strategiefond), ein eigenständiges Konzept weiterentwickeln und eine – wenn Sie so wollen – „Corporate identity" finden. Eine systemische Evaluation muß davon ausgehen, daß das Ganze mehr ist als die Summe seiner Teile. Das bisherige Evaluierungsverfahren der Leibniz-Institute erfaßt diese, nicht jedoch die Wissenschaftsgemeinschaft als Ganzes. Letzteres muß jedoch perspektivisch geschehen – und nicht nur mit Blick auf die Leibniz-Gemeinschaft –, um zu einer umfassenden Forschungsstrategie für Deutschland zu kommen, in welcher die Rollen sinnvoll verteilt und die Aufgaben klar formuliert sind. Auch eine systemische Evaluation kann nur dann gelingen, wenn die Aufgaben bekannt sind, auf die hin zu evaluieren ist.

Literaturverzeichnis

Internationale Kommission zur Systemevaluation der Deutschen Forschungsgemeinschaft und der Max-Planck-Gesellschaft (1999), Forschungsförderung in Deutschland, Hannover

Fischer, K., Evaluation der Evaluation, Teil I (1998), in: Wissenschaftsmanagement 5, S. 16-21

Fischer, K., Evaluation der Evaluation, Teil II (1998), in: Wissenschaftsmanagement 6, S. 17-23

Guba, E./ Lincoln, Y. (1989), Fourth Generation Evaluation, London

Heuer, H./Fuhlermann, H./Schmidt, K. H. (1998), Die Beurteilung von Forschungsleistungen, Frankfurt/Main

Lange, G. (1999), Keine Dienstmagd der Industrie, in: Frankfurter Rundschau 20, Mai 1999, S. 9

Liebald, C. (1996), Gutachten für die Vorstudie zur Evaluation der Weiterbildung, in: LWS Soest, Vorstudie zur Evaluation der Weiterbildung, Soest

Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz (Hg.) (1999), Stärke durch Vielfalt. Stellung und Bedeutung der Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz e.V. in der deutschen Forschungslandschaft, Bonn

Wissenschaftsrat (1996-1998), Stellungnahmen zu Instituten der Blauen Liste I-III, Köln

Martina Röbbecke

Einheitlichkeit oder Eigensinn?

Angemessene Indikatoren für heterogene Forschungseinrichtungen

Im Zentrum des Forschungsprojektes "Institutionelle Selbstbeobachtung als Steuerungsinstrument für außeruniversitäre Forschungseinrichtungen" stehen zwei Fragestellungen. Zum einen wird im Rahmen des Projektes untersucht, welche Bewertungskriterien dem breiten Spektrum der Einrichtungen der Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz (WGL) angemessen sind, und es werden Vorschläge für geeignete *Instrumente* der Evaluation von Blaue Liste-Instituten entwickelt. Zum anderen werden die Konturen eines *Verfahrens* erarbeitet, mit dem die Integration von unterschiedlichen *Zielen* beabsichtigt ist: es soll der externen Kontrolle dienen und die institutionellen Grundlagen der Selbststeuerung verbessern. Dieser Beitrag wird sich mit den Instrumenten der Forschungsevaluation auseinandersetzen, anschließend wird es um Fragen des Verfahrens und der Ziele von Evaluationen gehen.

Das erwähnte Projekt hat sich auf die Forschungseinrichtungen der WGL konzentriert, d.h. die zahlreichen Bibliotheken und Museen, die ebenfalls zur WGL gehören, wurden aufgrund ihrer besonderen Aufgaben nicht in das Untersuchungssample integriert. Aber auch die Forschungseinrichtungen der WGL sind ungewöhnlich heterogen. Daher ist jede Evaluation der WGL-Institute mit dem Problem konfrontiert, daß Bewertungskritierien und Indikatorenprogramme konstruiert werden müssen, die dem jeweils spezifischen Leistungsprofil sehr verschiedener Institute angemessen sind.

Grundsätzlich sind zwei Strategien denkbar, mit denen diese Aufgabe bewältigt werden kann. Eine Möglichkeit besteht darin, einen einheitlichen Satz von Indikatoren und Kriterien zu entwickeln, nach denen alle Einrichtungen begutachtet werden. Dann fiele – wie bei der aktuellen Evaluation durch den Wissenschaftsrat – den Gutachtern die zentrale Aufgabe zu, die Indikatoren zu gewichten und angemessen zu interpretieren. Es wäre aber auch möglich, ein Indikatorenprogramm zu konstruieren, das Modifikationen im Hinblick auf das jeweilige Leistungsprofil zuläßt, und daneben einen allgemeinen Katalog, der für alle Institute verbindlich ist und der zugleich eine vergleichende Bewertung ermöglicht. Diese Lösung würde auf eine Differenzierung zwischen einer "internen" und einer "externen" Evaluation hinauslaufen; selbstverständlich muß es eine Verbindung zwischen den beiden Katalogen geben. In jedem Fall bleibt allerdings die Spannung zwischen einer spezifischen, institutsbezogenen Bewertung und der routinisierten Anwendung eines standardisierten Verfahrens erhalten.

Den Verfahrensfragen soll hier nicht weiter nachgegangen werden, da sie, wie erwähnt, anschließenden ausführlich erörtert werden. An dieser Stelle wird es um die besonderen Probleme und Anforderungen einer institutsspezifischen Evaluation gehen, wobei auch

die Kommentare und Anregungen der von uns interviewten Mitarbeiter und Mitarbeiterinnen in fünf verschiedenen WGL-Instituten[15] berücksichtigt werden. Im Mittelpunkt der Darstellung stehen vier Thesen, die ich Ihnen vorstellen und anschließend erläutern möchte.

Zuvor jedoch eine Bemerkung zur Terminologie: Im folgenden wird nicht nur von "Indikatoren" gesprochen, sondern es soll zwischen Kennzahlen (beispielsweise die Zahl der vorhandenen Stellen), zwischen Prozeß- und Strukturmerkmalen (die im Vortragstitel noch als Throughput-Indikatoren bezeichnet werden) und Forschungsindikatoren im engeren Sinne unterschieden werden[16]. Diese Kriterien und ihre Funktion im Rahmen von Forschungsevaluationen der Blaue Liste-Institute sind Gegenstand der anschließenden Überlegungen.

These 1: Die Bewertungskriterien müssen angemessen im Hinblick auf die unterschiedlichen Aufgaben und Ziele der WGL-Einrichtungen sein. Um die Konzentration auf einen bestimmten Forschungstypus – in diesem Fall einen akademischen Bias – zu vermeiden, sollten für alle Leistungsbereiche jeweils adäquate Instrumente der Bewertung entwickelt werden.

Im Aufgaben- und Leistungsspektrum der WGL lassen sich vier Schwerpunkte unterscheiden. Die Forschungseinrichtungen bewegen sich auf sehr unterschiedlichen, überwiegend multi- oder interdisziplinär strukturierten Forschungsfeldern, sie sind durch eine Verbindung von grundlagen- und anwendungsorientierten Forschungsfragen gekennzeichnet, sie erbringen in einem erheblichen Umfang Beratungs- und Dienstleistungen und sie müssen, nicht zuletzt, Aufgaben bearbeiten, die von "überregionaler Bedeutung" und "in gesamtstaatlichem wissenschaftspolitischen Interesse" sind.

Je nach Institut ist die Bedeutung der einzelnen Aufgabenbereiche unterschiedlich. In einem Institut kann die interdisziplinäre Kooperation eine besondere Herausforderung darstellen, ein anderes ist darüber hinaus durch das breite Spektrum der Aufgaben zwischen Grundlagen- und anwendungsorientierter Forschung gekennzeichnet, während eine dritte Einrichtung nicht nur ein breites interdisziplinäres Forschungsspektrum, sondern auch die Spannung zwischen Politikberatung und eher akademisch orientierten Relevanzkriterien zu bewältigen hat.

Im Rahmen einer Evaluation müssen daher, so unsere Überlegung, zwei Aufgaben gelöst werden: es muß festgestellt werden, durch welches spezifisches Profil das zu begutachtende Institut gekennzeichnet ist, und es muß geklärt werden, welche Kriterien

[15] Zur methodischen Anlage des Forschungsprojektes vgl. auch das zwischenzeitlich erschienene WZB-Discussion Paper "Zwischen Reputation und Markt" mit ersten Auswertungen (Röbbecke/Simon 1999, S. 27 f.).
[16] Vgl. zu dieser Unterscheidung auch den folgenden Beitrag von Stefan Hornbostel.

die spezifischen Leistungen der jeweiligen Einrichtung erfassen und angemessen bewerten.

Bisher werden bekanntlich alle Forschungsinstitute auf der Grundlage eines einheitlichen Fragenkataloges begutachtet, der sich – neben Prozeß- und Strukturmerkmalen – überwiegend auf die Bewertung der *Forschungsleistungen* konzentriert und dabei die bekannten Wissenschaftsindikatoren einsetzt: beispielsweise werden Publikationen erfaßt, Drittmitteleinwerbungen und die Teilnahme an internationalen Fachtagungen.

In der kritischen Diskussion über die Verläßlichkeit und Gültigkeit von Wissenschaftsindikatoren wird unter anderem auf disziplinäre Unterschiede hingewiesen[17]. Sie führen beispielsweise dazu, daß die verschiedenen Publikationstypen einen unterschiedlichen Stellenwert haben. In einigen Disziplinen gelten Monographien und Handbücher als wichtige Publikationsformen, während in den natur- und technikwissenschaftlichen Bereichen eher Zeitschriftenartikel und in zunehmendem Umfang auch elektronische Publikationsformen verbreitet sind.

Nun kann man argumentieren, daß es durchaus möglich ist, diese Unterschiede bei der *Interpretation* der ermittelten Daten zu berücksichtigen. In der laufenden Evaluation ist allerdings ein akademischer Bias in der Bewertung durch die Gutachtergruppen des Wissenschaftsrates unverkennbar: Besonderer Nachdruck wird auf Veröffentlichungen in referierten Journals und auf erfolgreiche Drittmitteleinwerbung bei der Deutschen Forschungsgemeinschaft (DFG) gelegt.

Dieser Bias ist nicht nur für solche Institute nachteilig, deren Forschungsfelder durch eine andere Publikationspraxis gekennzeichnet sind. Generell ist problematisch, daß die Besonderheiten der Blaue Liste-Institute nicht angemessen berücksichtigt werden. Vor allem die erwähnte Verbindung von grundlagen- und anwendungsorientierter Forschung ist in der bundesdeutschen Forschungslandschaft unüblich: Während die MPG eher grundlagenorientierte und die FhG eher anwendungsorientierte Forschungsfragen untersucht, werden in vielen WGL-Einrichtungen sowohl Grundlagenforschung als auch Anwendungsforschung betrieben. Zum Teil sind in den WGL-Instituten theoretische Forschung und praktische Anwendung eng verbunden, die Ergebnisse der Grundlagenforschung können unmittelbar für die Lösung anwendungsnaher Probleme fruchtbar gemacht werden, und neue grundlagenorientierte Forschungsfragen werden im Kontext der Anwendung generiert. Die Verbindung, im besten Fall die Integration von Grundlagen- und Anwendungsforschung stellt ein spezifisches und zukunftsträchtiges Potential der WGL-Institute dar.

[17] Vgl. z.B. Daniel/Fisch (1988) und Hornbostel (1997).

Allerdings belohnt die Orientierung an Wissenschaftsindikatoren tendenziell erfolgreiche Leistungen der akademischen Wissenschaft. Wenn man eine Akademisierung der WGL-Institute verhindern will, die sie gerade ihrer Besonderheiten berauben und zu einer stärkeren Konkurrenz mit anderen Institutionen führen würde, halten wir es für unverzichtbar, alle Leistungsbereiche in die Bewertung einzubeziehen. Wir versuchen daher, Kriterien zu entwickeln, mit denen auch der Aufgabenbereich anwendungsorientierter Forschung sowie die Beratungs- und Dienstleistungsaufgaben berücksichtigt werden können. Dazu gehören beispielsweise die Berücksichtigung von Interessen der Adressaten und der Aufbau von Kooperationsstrukturen mit potentiellen Nutzern. Grundsätzlich stellt sich allerdings das Problem, daß viele Leistungen kaum durch quantitative Indikatoren zu erfassen sind.

Wir schlagen vor, zukünftig die Nutzer und Adressaten der Forschungs- und Beratungsaufgaben stärker in das Begutachtungsverfahren zu integrieren, da sie bei der Entwicklung ergänzender Leistungskriterien hilfreich sein können. Auf der anderen Seite sind aber auch die Institute gefordert, ihre Beratungsleistungen zu professionalisieren, ihre umsetzungsrelevanten Forschungsergebnisse besser sichtbar zu machen und präzise Zielvorstellungen zu entwickeln. Zur Aufgabe von Evaluationen gehört es dann auch, diese Ziele und die Wege zu ihrer Realisierung zu begutachten.

These 2: **In vielen Einrichtungen gibt es Vorbehalte gegenüber den bisher verwendeten Indikatoren und Bewertungskriterien. Insbesondere wird kritisiert, daß teilweise ein Interpretationsrahmen für die erhobenen Daten fehle oder aber – ganz im Gegenteil – manche Vorgaben zu unflexibel seien. Für eine Fortsetzung der begonnenen Debatten über das eigene Selbstverständnis und zukünftige Profil ist es wichtig, daß die Einrichtungen die Indikatoren und Bewertungskriterien akzeptieren und sie weiterentwickeln können.**

Ein Hintergrund der genannten Kritik ist in dem erheblichen administrativen Aufwand zu sehen, der betrieben werden muß, um den umfangreichen Fragenkatalog des Wissenschaftsrates zu beantworten. Beispielsweise sind die Fragen zur personellen Ausstattung ausgesprochen detailliert – so werden neben der Stellenausstattung auch die Dienstbezeichnungen, das Alter und das Eintrittsjahr, Geschlecht, Ausbildungsabschluß und Eingruppierung erfragt.[18]

Sicherlich wird sich der Aufwand, der zur Ermittlung dieser Daten notwendig ist, in den nächsten Jahren verringern lassen, wenn den zukünftigen Erhebungen zumindest teilweise die heutigen Datensätze zugrunde gelegt werden können. Gleichwohl ist fraglich, ob der große Erhebungsaufwand zu interpretationsfähigen Daten geführt hat und, noch grundsätzlicher, ob und welche Auskünfte diese Daten über die Leistungsfähigkeit einer Forschungseinrichtung geben können.

[18] Vgl. Wissenschaftsrat (1997).

Während bei den soeben genannten Daten unsicher ist, in welcher Weise sie interpretiert werden, gibt es hinsichtlich des Verhältnisses zwischen der Zahl der unbefristet und der befristet tätigen MitarbeiterInnen klare Vorgaben, die sich aus Empfehlungen des Wissenschaftsrates aus dem Jahr 1993 ableiten: er erwartet, daß zwischen 30% und 50% der Stellen befristet besetzt werden.[19] Gegen diese Quote regt sich in den Instituten Widerspruch; die Vorgabe wird häufig als zu starr bezeichnet und nicht akzeptiert, da sie mit den Besonderheiten der Einrichtung nicht kompatibel sei. Insbesondere wird kritisiert, daß es im Rahmen des Bewertungsverfahrens kaum möglich sei, die jeweilige Befristungsstrategie zu erläutern und zu kommentieren.

Ähnlich zurückhaltend betrachten viele InterviewpartnerInnen die Fragen nach den eingeworbenen Drittmitteln, bei denen auf der Abteilungsebene detailliert die Quellen, das jeweilige finanzielle Volumen und die Laufzeit der Drittmittelprojekte erfragt werden.

Vielen Instituten war anfangs nicht bekannt, vor welchem Hintergrund diese Daten interpretiert werden. Die Befürchtung einer rein quantifizierenden Bewertung – bei der die Höhe der insgesamt eingeworbenen Drittmittel relevant ist – erwies sich als unbegründet, dagegen zeigt sich im Laufe des derzeitigen Evaluationsverfahrens, daß die Drittmittelakquise bei einem bestimmten Drittmittelgeber besonders honoriert wird. Die Gutachtergruppen des Wissenschaftsrates legen erheblichen Wert auf die Einwerbung von Drittmitteln bei der DFG. Gegen diese Wertung richtet sich vielfacher Einspruch, auf den hier nur kurz eingegangen werden kann. Zentraler Einwand ist, daß die überwiegend grundlagenorientierten Forschungsprojekte, die von der DFG gefördert werden, nur einen kleinen Teil des Forschungsspektrums abbilden, das von den WGL-Instituten bearbeitet wird. Außerdem haben die Einrichtungen erhebliche Probleme, interdisziplinäre Forschungsprojekte von der nach disziplinären Gesichtspunkten organisierten DFG bewilligt zu bekommen, und sie kritisieren den hohen Aufwand der Antragserstellung für Mittel, die die Gesamtkosten eines Projektes in der Regel nicht abdecken. Generell beanstanden die Institute, daß sie, ähnlich wie bei der Befristungsproblematik, kaum Möglichkeiten haben, ihre jeweiligen Drittmittelstrategien – die von Abteilung zu Abteilung durchaus differieren können – darzustellen.

Die hier skizzierte Problematik umfangreich erhobener Daten, die erst interpretiert werden müssen, und die Beeinflussung der Interpretation durch wissenschaftspolitische Argumente oder durch die universitäre Herkunft der Gutachter ist bisher vor allem unter dem Gesichtspunkt betrachtet worden, daß darunter die Akzeptanz von Evaluationen leidet. Wichtig ist aber auch, daß starre Bewertungskriterien dem Handlungsspielraum der Einrichtungen Grenzen setzen. Wir halten es daher für wichtig, daß sich die Datenerhebung auf aussagefähige Daten konzentriert und daß im Zusammenhang mit der Datenerhebung stärker als bisher auch die Ziele der Institute erfragt werden.

[19] Vgl. Wissenschaftsrat (1994).

Dadurch haben die Institute eine Möglichkeit, ihre Strategien und Entwicklungsoptionen vorzustellen und zugleich inhaltliche Kommentierungen abzugeben, die voreilige Schlußfolgerungen und Fehlinterpretationen seitens der Gutachter verhindern können.

These 3: Die Verwendung von Forschungsindikatoren kann mit Nebeneffekten verbunden sein, die nicht in jedem Fall erwünscht sind. Es gilt daher, diese Effekte zu identifizieren und ihnen gegebenenfalls entgegenzuwirken.

Es dürfte bereits deutlich geworden sein, daß die in einem Evaluationsverfahren verwendeten Forschungsindikatoren nicht nur bewerten, sondern auch eine steuernde Wirkung entfalten. Selbstredend bereiten sich die Forschungseinrichtungen auf eine Evaluation vor und versuchen, den Ausgang zu ihren Gunsten zu gestalten. Die Anstrengungen und Strategien konzentrieren sich dabei auf jene Indikatoren, deren hohe Relevanz innerhalb der scientific community unbestritten zu sein scheint oder deren Bedeutung innerhalb der mehrjährigen Evaluation der WGL-Institute deutlich geworden ist.

Diese steuernde Funktion ist beabsichtigt, sie kann allerdings zu unerwünschten oder sogar schädlichen Ergebnissen führen, wie anhand von zwei Beispielen gezeigt werden soll.

Im Rahmen der derzeitigen Evaluation spielen *Publikationen in referierten Zeitschriften* eine herausgehobene Rolle. Es wurde bereits erwähnt, daß es fachspezifische Unterschiede in der Publikationspraxis gibt – so ist den Naturwissenschaften der Aufsatz der dominierende Publikationstypus, aber nicht in den Sozialwissenschaften. Noch schwerwiegender ist, daß durch die normative Setzung eines bestimmten Publikationstypus die Publikationsstrategien der Institute beeinflußt werden – allerdings in einer Weise, durch die ihre Besonder- und Eigenheiten nivelliert werden könnten.

Bekanntlich wird befürchtet, daß der Versuch, die Publikationschancen in referierten Zeitschriften zu optimieren, zu einer wachsenden Orientierung am Mainstream der Forschungsfragen führt. Diese Gefahr verstärkt sich, wenn die Bedeutung der Zeitschriften, Monographien und Sammelbände, die ein Institut herausgibt, nicht angemessen bewertet werden. Die Veröffentlichung von Forschungsergebnissen in diesen Publikationen als "in-house-Publikationen" zu bewerten, wie es zum Teil geschehen ist, verkennt den forschungsstrategischen Stellenwert, den sogenannte "Hauszeitschriften" in Spezialgebieten haben können. Die grundsätzlich niedrigere Bewertung von Sammelbänden widerspricht wiederum der gezielten Förderung von internationalen Forschungskooperationen, deren Ergebnisse sich nun einmal am besten in dieser Form veröffentlichen lassen.

Ferner führt die Orientierung an referierten Zeitschriften, die oftmals englischsprachig sind und auf internationaler Ebene erscheinen, zu wachsenden Schwierigkeiten,

spezifisch nationale oder gar regionale Forschungsthemen dort zu verankern. Andererseits könnte es jedoch ausgesprochen problematisch werden, wenn die Institute diese Fragen, die in wissenschaftspolitischer Sicht wichtig sind, aus den Augen verlieren. Nicht zuletzt kann die Orientierung an referierten Journals dazu führen, daß eher anwendungsorientierte Einrichtungen und ihre stärker adressatenbezogene Publikationspolitik nicht angemessen bewertet werden.

Als zweites Beispiel soll noch einmal kurz auf die bereits erwähnte Drittmittelakquise eingegangen werden. Die hohe Bewertung von DFG-Mitteln verstärkt die Orientierung auf Fragestellungen der Grundlagenforschung, die jedoch nur einen Teilbereich des Forschungsspektrums der WGL-Institute darstellt. Diese Steuerung wird in den letzten Jahren durch erhebliche finanzielle Sanktionen und Anreize verstärkt. Die WGL-Institute sind dazu verpflichtet worden, 2,5% ihrer Grundfinanzierung an die DFG abzuführen. Im Gegenzug können sich die Institute um eine Förderung durch die DFG auch auf ihren Hauptarbeitsgebieten bemühen – das ist neu. Dadurch haben einige Institute erstmals Zugang zu DFG-Mitteln. Die Institute haben also ein weiteres wichtiges Motiv, sich um DFG-Mittel zu bewerben, die damit verbundene Konzentration auf eher grundlagenorientierte Forschungsfragen dürfte jedoch ihrer Profilierung nur begrenzt nützlich sein. Bei dieser "Übersteuerung" handelt es sich um ein ungeplantes Resultat, um eine zufällige Koinzidenz. Um ähnliche Effekte zukünftig zu vermeiden, sollte die derzeitige Evaluation sorgfältig ausgewertet und nach Formen gesucht werden, die eine kritische Begleitung und Kommentierung von Evaluationen ermöglichen.

These 4: Evaluationen von wissenschaftlichen Einrichtungen haben auch deren Struktur- und Prozeßmerkmale zum Gegenstand. Dabei muß allerdings berücksichtigt werden, daß es für die heterogenen Institute kein einheitliches Organisationsmodell geben kann. Entscheidendes Kriterium ist, ob die vorgefundenen Organisationsvarianten dem jeweils spezifischen Leistungsprofil entsprechen.

Eine Besonderheit der derzeitigen Evaluation durch den Wissenschaftsrat besteht in der hohen Bedeutung, die den Organisationsstrukturen der Einrichtungen beigemessen wird. Nicht nur die "Produkte" der Forschungsarbeiten werden begutachtet, sondern auch wichtige Merkmale der Aufbau- und Ablauforganisation. Grundsätzlich begrüßen wir diese Herangehensweise, da die Organisationsstruktur eine entscheidende Voraussetzung für die Leistungsfähigkeit einer Forschungseinrichtung darstellt. Normative Setzungen sollten jedoch mit großer Vorsicht vorgenommen werden, da sie zu Fehlentwicklungen und Akzeptanzproblemen führen können.

Zwei Beispiele sollen im folgenden diese Zurückhaltung begründen. Der Wissenschaftsrat fragt im derzeitigen Evaluationsverfahren nach den abteilungsübergreifenden Forschungsprojekten und möchte wissen, welche Organisationseinheiten daran beteiligt sind. Durch diese Frageweise wird, wie wir aus den

Interviews wissen, der Eindruck erweckt, daß abteilungsübergreifende Kooperationen ein Indikator für interdisziplinäre Kooperation oder gar ein ein Selbstzweck seien. Auch in forschungspolitischen Diskussionen ist die abteilungsübergreifende Kooperation ein modisches Konzept, mit dem Synergieeffekte versprochen werden. Dagegen ist allerdings einzuwenden, daß eine hohe Zahl von abteilungsübergreifenden Kooperationen auch ein Indikator dafür sein kann, daß das Prinzip der Abteilungsgliederung der Aufgabenstruktur des Instituts nicht angemessen ist. Wenn beispielsweise eine Einrichtung interdisziplinäre Forschungsprojekte bearbeitet, die Abteilungen jedoch disziplinär strukturiert sind, ist eine abteilungsübergreifende Zusammenarbeit zwingend notwendig. Es kann gute Gründe für diese Organisationsstruktur geben, allerdings ist sie mit einem hohen Planungs- und Steuerungsaufwand verbunden. Die Alternative ist, die Abteilungen interdisziplinär zu strukturieren und folglich die Häufigkeit von abteilungsübergreifenden Kooperationen erheblich zu senken.

Das zweite Beispiel bezieht sich auf ein Institut, das positiv evaluiert worden ist. Trotz des guten Ergebnisses haben die Gutachter Veränderungen der Organisationsstruktur empfohlen: sie erwarten, daß die – im Vergleich mit anderen Einrichtungen – große Zahl der Abteilungen verkleinert wird. Eine höhere Leistungsfähigkeit der neuen Strukturen kann jedoch nicht zweifelsfrei begründet werden – im Gegenteil ist anzunehmen, daß die kleinere Abteilungszahl bei einer insgesamt gleichbeibenden Institutsgröße zu einer Formalisierung und Zentralisierung der Organisationsstruktur führen wird. Die Empfehlungen der Gutachter beruhen offensichtlich auf einem Organisationsmodell, das in einem anderen Kontext durchaus erfolgreich sein mag, aber keineswegs für alle WGL-Institute optimal ist.

Über formale Strukturmerkmale hinaus sollten in der zukünftigen Evaluation der WGL auch informelle Strukturen und Prozeßmerkmale stärker beobachtet werden. Die Analyse der formalen Strukturen und das Studium der Satzungen erlaubt es in der Regel kaum, Informationen über darüber zu gewinnen, in welcher Weise zentrale Aufgaben einer Forschungsorganisation gelöst werden: also wie beispielsweise Projekte generiert werden, nach welchen Kriterien Projekte fortgeführt oder beendet und auf welcher Grundlage die Ressourcen verteilt werden.

Abschließend soll zusammenfassend festgehalten werden:

- um zu gewährleisten, daß die Bewertungskriterien den jeweils spezifischen Institutsaufgaben und -zielen angemessen sind, sollten die Institute zukünftig stärker an der Konzeption der Evaluation beteiligt werden,
- wichtiger als die Konzentration auf Forschungsindikatoren ist die Entwicklung eines Verfahrens, das eine Stärke-Schwächen-Analyse des gesamten Leistungsspektrums erlaubt,

- die Organisationsanalyse sollte durch den Einbezug von Experten professionalisiert und die Gutachter gezielt auf ihre Aufgaben vorbereitet werden,
- und schließlich sollten die Empfehlungen durch wissenschafts- und organisationssoziologische Studien untermauert werden.

Literaturverzeichnis

Daniel, H.-D. / Fisch, R. (Hg.) (1988), Evaluation von Forschung. Methoden, Ergebnisse, Stellungnahmen. Konstanzer Beiträge zur sozialwissenschaftlichen Forschung Bd. 4, Konstanz

Hornbostel, S. (1997), Wissenschaftsindikatoren, Opladen

Röbbecke, M. / Simon, D. (1999), Zwischen Reputation und Markt. Ziele, Verfahren und Instrumente von (Selbst)Evaluationen außeruniversitärer, öffentlicher Forschungseinrichtungen, WZB-Discussion Paper (P 99-002), Berlin

Wissenschaftsrat (1994), Empfehlungen zur Neuordnung der Blauen Liste, in: ders., Empfehlungen und Stellungnahmen 1993, S. 453 ff.

Wissenschaftsrat (1997), Fragebogen für die Bewertung der Forschungseinrichtungen und Museen der Blauen Liste (Drs. 2888/97), Köln

Stefan Hornbostel

Welche Indikatoren zu welchem Zweck: Input, Throughput, Output

In den sechziger und siebziger Jahren erregte die Frage, ob man Forschungsleistungen messen könne, noch einiges Aufsehen. Obwohl Wissenschaft und Forschung auch heute mit Begriffen wie Unwägbarkeit, Zufall, Nicht-Planbarkeit, Nicht-Messbarkeit etc. assoziiert werden, wirkt die Frage heute weniger provokant, denn inzwischen haben wir uns daran gewöhnt, daß fast alles wenn nicht exakt vermessen, so doch öffentlicher Evaluation zugeführt wird. Dabei spielen Indikatoren, die ursprünglich meist in einem analytischen Kontext entwickelt wurden, eine immer wichtigere Rolle. Zum einen stoßen die klassischen Verfahren des Peer Review angesichts weiter steigenden Bedarfs an Evaluation, Monitoring, öffentlicher Rechenschaftslegung und öffentlichem Leistungsvergleich an ihre Grenzen. Kompetente Peers stehen nur in begrenzter Zahl zur Verfügung und sind auch nur begrenzt belastbar. Zudem sind derartige Begutachtungen sehr aufwendig und für Routineberichterstattungen kaum geeignet. Schließlich geht es zunehmend um den Vergleich von Institutionen, um „best practice"-Modelle und nicht so sehr um die Erfüllung von Mindeststandards. Dazu aber sind meist Vergleiche notwendig, die eine Quantifizierung von Leistungen und eingesetzten Ressourcen erfordern. Last not least spielen Indikatoren eine immer größere Rolle bei der Verteilung von (zunehmend globalisierten) Haushaltsmitteln.

Indikatoren sollen Repräsentanten für eine meist wenig überschaubare und hochkomplexe Realität sein. Und sie sollen diese Realität in wenigen Zahlenwerten repräsentieren. Allerdings ist die zu indizierende Realität – nicht nur im Falle Wissenschaft – meist nur sehr vage gekennzeichnet: Schwammige Begriffe wie Armut, wirtschaftliche Gesundheit, Fortschritt, gute Forschung etc. stehen bei Wirtschafts-, Sozial-, aber auch Wissenschaftsindikatoren im Hintergrund. Voraussetzung ist aber, daß diese Begriffe soweit operationalisiert und modelliert werden können, daß sich ein Indikator konstruieren läßt. Das unterscheidet Indikatoren von Kennzahlen, hinter ersteren stehen eine modellhafte Annahme über die Realität und ein Operationalisierungskonzept.

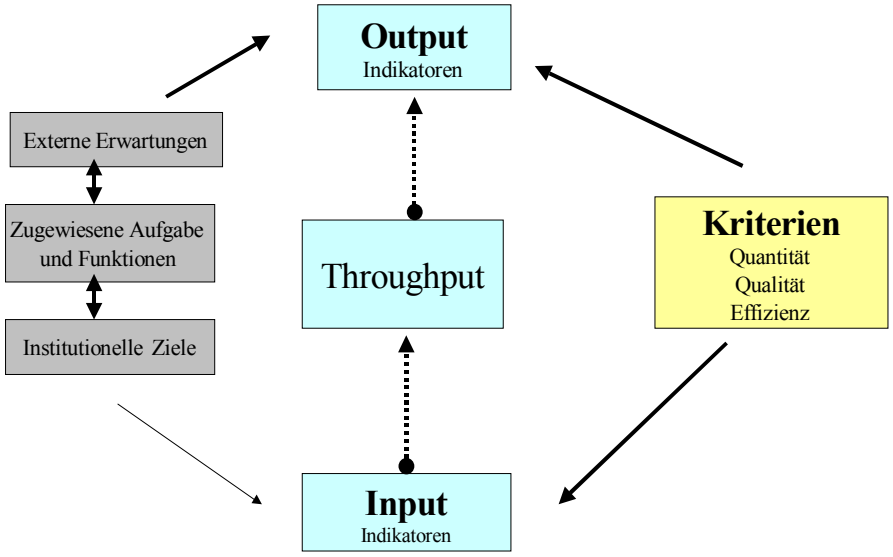
Allein im Rahmen akademischer Betrachtungen waren Wissenschaftsindikatoren eher eine kuriose, jedenfalls undramatische Angelegenheit. Werden Indikatoren aber über diesen Rahmen hinaus bedeutsam, dann sind Interessen tangiert und die Angemessenheit von Indikatoren steht zur Debatte. Daß die Auswahl und Konstruktion von Indikatoren bestimmte Akteure bevorzugt bzw. zurücksetzt, ist kein Defekt der Indikatoren, sondern lediglich ein Verweis darauf, daß Indikatoren nicht per se eine Wirklichkeit abbilden. Indikatoren antworten auf Fragen, und sie tun dies meist mit allerlei Unzulänglichkeiten. Ähnliches gilt für die Frage der Steuerungswirkung von Indikatoren. Ob Indikatoren gewünschte Effekte erzeugen, akzeptable Kollateralschäden produzieren oder kontraproduktiv wirken, läßt sich nur anhand einer definierten Zielvorstellung abschätzen.

Indikatoren: Bindung an Ziele

Wenn es um die Evaluation von Institutionen geht oder um Informationsgrundlagen für Steuerungsprozesse, dann heißt das, daß Indikatoren an Zielvorstellungen und Aufgabendefinitionen einerseits und Fragestellungen andererseits rückgebunden sind. Am Beispiel eines ökonomischen Indikators kann man sich diese Zielgebundenheit sehr deutlich machen. Die Berechnung des Bruttosozialprodukts stellt wesentlich auf Wachstumsprozesse im Rahmen rechtlich organisierter Wirtschaft ab. Qualifiziert man nun die Art des wirtschaftlichen Handelns, nimmt man beispielsweise eine ökologische Position ein, dann erscheinen alle Beiträge zur Beseitigung von Umweltschäden nicht mehr als Wachstumsgrößen, sondern mit negativem Vorzeichen als Reparaturleistung. Anders formuliert: Der Indikator „BSP“ wäre einer Zielformulierung „Beschreibung substantiellen Wachstums“ nicht angemessen.

Durchaus ähnlich sind die Verhältnisse im Bereich von Forschung und Entwicklung. Zunächst entsteht die Frage nach den Zielen oder Aufgaben einer Institution: Bestehen sie primär in Grundlagenforschung, stärker in der Entwicklung anwendungsbezogener Problemlösungen, in Beratungsaufgaben, in der Organisation von Transferprozessen, in der Bereitstellung von Infrastruktur usw.? Derartige Fragen sind besonders dann wichtig, wenn ein sehr heterogenes Institutionenensemble wie etwa die „Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz“ evaluiert werden soll.

Abbildung 1:



Die Zieldefinition ist mit einem gesetzlichen Auftrag meist nur unzureichend beschrieben. Die Interpretation eines solchen Auftrags durch die Institution und die externen Erwartungshaltungen müssen keineswegs völlig kompatibel sein. Teil eines Evaluationsprozesses wird daher immer auch die genaue Klärung der Ziele und angestrebten Entwicklungen (mission) sein müssen. Das fällt bei einem hoch spezialisierten Forschungsinstitut leichter als bei einer Einrichtung, in der Service, Forschungs- und Entwicklungsaufgaben gleichzeitig erfüllt werden sollen.

Kriterien der Leistungsbeurteilung

Sodann stellt sich die Frage nach Beurteilungskriterien: Was ist wesentlich? Herausragende Forschungsleistungen, effiziente Mittelverwendung, breite Präsenz in einem Fachgebiet, interdisziplinäre Kontakte, internationale Kooperationen, Mitarbeit an Forschungsfronten oder zentralen fachlichen Themen, Kundenzufriedenheit, Erfüllung definierter Qualitätsstandards? Die Liste läßt sich fast beliebig erweitern. Die einzelnen Punkte schließen sich auch nicht aus, aber sie setzen Akzente, die am Ende jeweils unterschiedliche Indikatorenkonstrukte nahelegen. Was ein geeigneter Indikator ist, bestimmt sich also zunächst einmal durch die Definition einer Leistungsdimension und dann durch eine Spezifikation des Leistungsbegriffs.

Verwendungsmöglichkeiten

Zu welchem Zweck können nun Indikatoren eingesetzt werden? Indikatoren sind meist in sehr unterschiedlichen Kontexten verwendbar. Einschränkungen ergeben sich weniger hinsichtlich des Verwendungszwecks als vielmehr im Hinblick auf die disziplinäre Eignung, ihre Fehleranfälligkeit, die Aggregatebene, auf der einzelne Indikatoren verwendbar sind, und hinsichtlich der Aktualität.

Zumindest drei unterschiedliche Verwendungskontexte sollten aber unterschieden werden:

Der erste betrifft das Verhältnis von Institution und Öffentlichkeit oder öffentlichen Finanziers. Hier werden Indikatoren für eine Art Rechenschaftslegung eingesetzt. Geht die Initiative von der Institution aus, ist sie relativ frei in der Bestimmung der Indikatoren, kommt die Initiative von außen, wird sie nolens-volens auf bestimmte Informationsanforderungen reagieren müssen.

Das Gegenstück wäre ein internes „Monitoring“. Hier ist man weitgehend frei von externen Erwartungen. Die Fragen, die sich auf dieser Ebene stellen, beziehen sich eher darauf, wie ein aussagefähiges Controling system mit vertretbarem Erhebungsaufwand aussehen kann. Es geht dort nicht um einmalige großangelegte Evaluationen, sondern um eine kontinuierliche indikatorengestützte Selbstbeobachtung.

Wenn man auf die Universitäten, aber auch auf die Allokationspolitiken mancher Länder schaut, wird ein dritter Kontext deutlich, nämlich die Verbindung von Leistungs-, Belastungs- oder Performanceindikatoren mit Mittel- oder Personalzuweisungen. Hier zeigt sich eine andere Eigenschaft von Indikatoren. Sie sind nämlich – einmal etabliert – so etwas wie institutionalisierte Dritte. Sie entlasten Zuweisungsmechanismen von Dauerverhandlungen. Sie etablieren sozusagen eine eigene Faktizität. Zugleich wird damit allerdings auch voluntaristisches Umsteuern schwieriger.

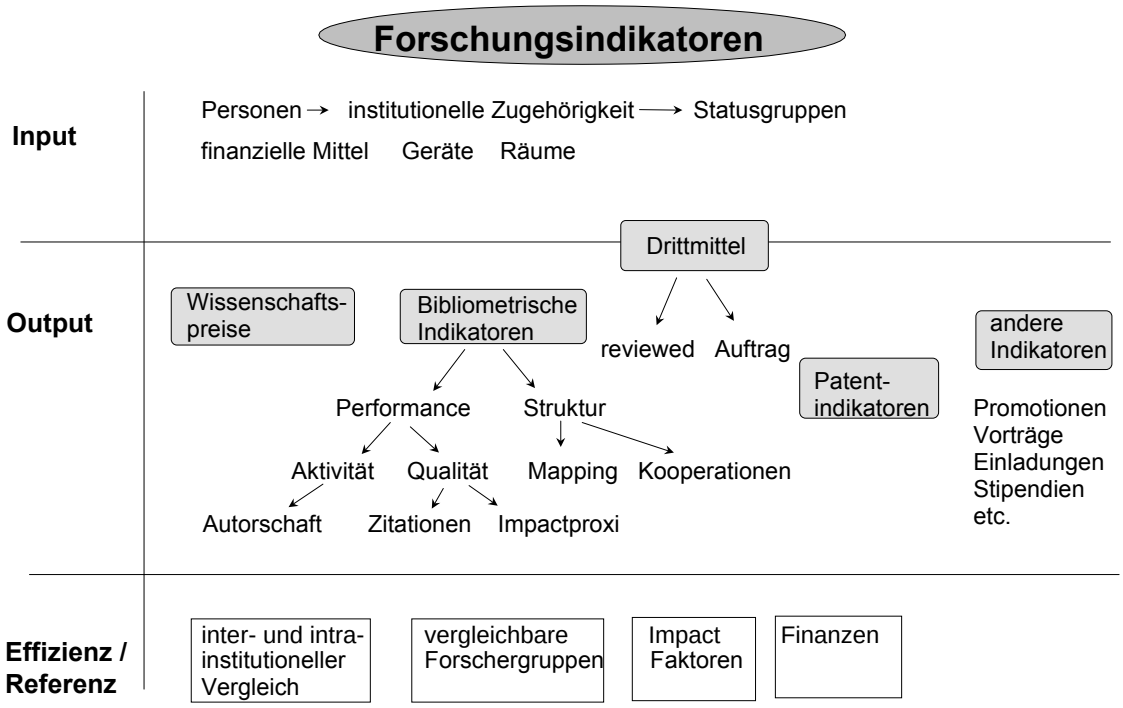
Welche Indikatoren?

Traditionell hat man in Deutschland den Forschungsbereich vor allen Dingen mit Input-Indikatoren beschrieben (vgl. etwa die Forschungsberichte der Bundesregierung). In den letzten Jahren haben sich aber immer stärker Output-Indikatoren durchgesetzt. Ein Trend, der sich in vielen gesellschaftlichen Bereichen abzeichnet und mit dem Begriff „Audit Society“ durchaus treffend beschrieben ist.

Input-Indikatoren spielen aber insofern weiterhin eine wichtige Rolle, als sie zur Beurteilung von Aufwand/Ertrag-Relationen und zum Vergleich der „Effizienz“ von Institutionen notwendig sind. Was jeweils genau zum Input zu rechnen ist bzw. welche Anteile von Ressourcen zu berücksichtigen sind, ist in der Praxis häufig gar nicht so einfach zu bestimmen und wird nicht selten über Schätzungen ermittelt (z.B. Vollzeitäquivalente des Personals für FuE an den Hochschulen). Diese Unsicherheitszonen machen insbesondere Effizienzvergleiche anfällig für Fehleinschätzungen.

Schwieriger zu ermitteln ist der Output. Zunächst muß dazu das „Produkt“ (s. o.) näher bestimmt und gegebenenfalls abgegrenzt werden (Forschung, Service, Lehre, Nachwuchsförderung etc.), sodann geeignete Indikatoren zur Messung dieses Outputs bestimmt werden. Das geht nicht ohne ein gewisses Maß an konsensualer Festlegung von Konventionen (wie er in paradigmatisch verfestigten Forschungsgebieten – z. B. in der Physik – durchaus existiert). Der ganz überwiegende Teil der Forschungsindikatoren greift dazu Bewertungsprozesse auf, die innerhalb der Wissenschaft ohnehin anfallen, und verdichtet sie zu numerischen Charakterisierungen. Derartige Bewertungen fallen in der Manuskriptbeurteilung, in der Beurteilung von Forschungsförderungsanträgen, in der Rezeption wissenschaftlicher Erkenntnisse und ihrer Dokumentation (Zitat), in der organisierten Honorierung von Leistungen (Preise, Ehrungen) und vielen anderen Formen der Begutachtung statt. Einige wenige Indikatoren versuchen auch die Wirkungen von Forschung in anderen gesellschaftlichen Subsystemen (Wirtschaft, Kultur) meßbar zu machen.

Abbildung 2:



S. Hornbostel: Inst. für Soziologie der FSU Jena

Die Abgrenzung zwischen Input- und Output-Indikator ist gelegentlich problematisch, denn die Interpretation eines Indikators kann sich durchaus mit dem Focus der Analyse ändern. Aus der Perspektive eines Drittmittelgebers sind z. B. Drittmittelbewilligungen zunächst einmal Input für die geförderte Institution/Person, der Output ist das, was am Ende aus einem Forschungsprojekt hervorgegangen ist. Aus der Perspektive eines Forschers oder eines Instituts ist die Einwerbung von Mitteln aber bereits das Ergebnis einer durchaus aufwendigen Antragsarbeit, mit der man sich gegen Konkurrenten durchgesetzt hat, in diesem Sinne also ein Output. Ähnliches gilt für das Personal, wenn man Qualifizierungsprozesse im Auge hat: Es handelt sich mal um Input, mal um Output.

Mit der Verrechnung von Input- und Outputgrößen bzw. mit dem Vergleich von Outputangaben einer Institution an geeigneten und Referenzgrößen lassen sich schließlich Effizienzmaße konstruieren bzw. eine Positionierung relativ zu Vergleichseinheiten vornehmen.

Was messen Indikatoren?

Grob lassen sich Indikatoren danach unterscheiden, ob sie Inputgrößen beschreiben, Aktivität oder Partizipation messen, Qualität, Sichtbarkeit oder Akzeptanz wiedergeben, Strukturen und Prozesse abbilden oder aber subjektive Einschätzungen.

Aktivitätsindikatoren wie etwa Publikationszählungen, Drittmittelanträge, Patentanmeldungen berichten zunächst nur über die Präsenz im wissenschaftlichen Kommunikationssystem bzw. die Teilhabe an Transferprozessen in das ökonomische System.

Qualitätsindikatoren, z. B. Zitationsanalysen, Lizenzanalysen, Bewilligungen von Drittmitteln greifen auf Bewertungsprozesse in der Scientific Community zurück oder auf Bewertungen von Nutzern von Forschungsergebnissen. Dabei geht es meist weniger um ein methodologisches Qualitätskonzept als um Resonanz, Akzeptanz bestimmter Befunde oder Anknüpfung/Fortführung von Forschungslinien.

Strukturindikatoren knüpfen nicht so sehr an Leistungsdimensionen an, sondern beschreiben spezifische Eigenschaften des Output oder aber des Erstellungsprozesses. Sie bemühen sich um Deskription der Forschungstätigkeiten einer Untersuchungseinheit im Hinblick auf thematische Verortung, auf Kooperationsstrukturen, Interdisziplinarität, die Teilnahme an aktuellen Forschungsfronten etc.

Auf subjektive Einschätzungen wird man vor allen Dingen dann zurückgreifen müssen, wenn man mit klassischen Indikatorenkonzepten nicht weiter kommt. So wird man z. B. Forschungsmuseen nur teilweise über typische Forschungsindikatoren beschreiben können. Darüber hinaus würde es in einem solchen Fall Sinn machen, sich an Modellen zu orientieren, die im kommunalen Bereich beim Leistungsvergleich von Kultureinrichtungen entwickelt wurden, dazu gehören dann z. B. Nutzerbefragungen, Mitarbeiterbefragungen, Sponsoringanalysen etc.

Fallstricke und Fettnäpfchen

Beim Einsatz solcher Indikatoren sind allerdings immer auf mehreren Stufen Entscheidungen zu treffen, die Einfluß darauf haben, welcher Realitätsausschnitt modelliert wird. Zunächst ist zu klären, welche Leistungen ein Indikator erfaßt bzw. erfassen soll. Bereits auf dieser Stufe sind häufig Gewichtungen und Bewertungen notwendig, deren Gelingen entscheidend davon abhängig ist, inwieweit disziplinäre Konventionen über Wertigkeiten von Forschungsergebnissen vorhanden sind.

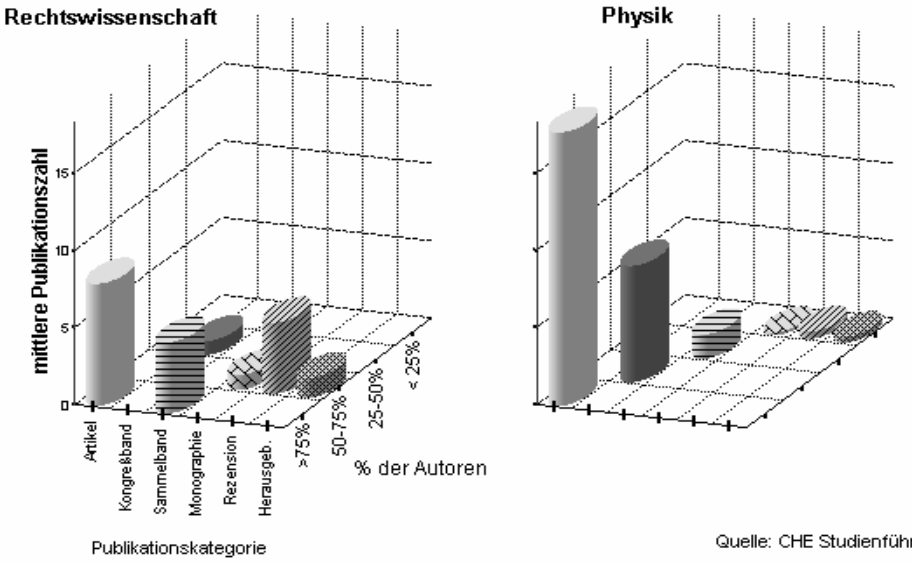
Am Beispiel von Publikationsanalysen und Drittmitteleinwerbungen seien die einzelnen Entscheidungsschritte und Konsequenzen kurz dargestellt:

Publikationen und Zitate

Zunächst muß entschieden werden, welche Publikationstypen berücksichtigt werden sollen. In der Physik fällt eine solche Entscheidung relativ leicht. Wie die folgende Abbildung zeigt, ist mit Artikeln in Fachzeitschriften und Kongreßbeiträgen der ganz überwiegende Teil der Publikationen erfaßt. Zudem ist die Reputation der verschiedenen Journals nicht sonderlich umstritten, so daß z. B. mit dem Science Citation Index oder spezialisierteren Fachdatenbanken in der Regel problemlos gearbeitet werden kann. Ganz anders die Situation bei den Rechtswissenschaftlern: Dort spielen Monographien und als Reflex darauf Rezensionen eine wichtige Rolle. Artikel in Fachzeitschriften streuen über die unterschiedlichsten journals (auch nicht-juristische). Die Unterscheidung zwischen originärer Forschung, Kommentierung und Dokumentation ist schwer zu ziehen, und über die Wertigkeit einzelner Zeitschriften herrscht wenig Einverständnis. Wollte man die juristischen Publikationen zu einem Indikator verarbeiten, wäre es nicht nur notwendig, die unterschiedlichen Publikationstypen auf irgendeine Weise zu verrechnen, sondern auch den diversen Zeitschriften Gewichte zuzuordnen.

Abbildung 3:

Durchschnittliche Publikationszahlen nach Publikationskategorie und Prozentsatz der Befragten, die mind. eine Publikation aus der jeweiligen Kategorie angegeben haben. Zeitraum: 1995 bis 1997



Quelle: CHE Studienführer 1999

Ist eine Liste von Publikationen erstellt, müssen weitere Entscheidungen getroffen werden. Die wichtigste ist dabei die Zuordnung von Publikationen zu den Untersuchungseinheiten: Das können Personen, Forschergruppen, Labore, Abteilungen, Institute, Universitäten, nationale Forschungssysteme oder auch weltweite Entwicklungen sein.

Sobald es um kleinere Aggregate geht, wird die Zuordnung von Texten zu Personen oder Institutionen in dem Augenblick schwierig, wo es sich um mehrere Autoren handelt. In den Naturwissenschaften ist das inzwischen die Regel, in Extremfällen tauchen mehr als 100 Autoren auf. Eine einfache Zählung führt dann zu Ergebnissen, wonach die Spitzenschreiber etwa alle vier Tage eine Publikation veröffentlichen. Der Grund liegt in institutionell durchaus unterschiedlichen Konventionen dafür, wer als Autor erscheint: Ob der Laborchef auf jeder Publikation als Autor geführt wird, ob die technischen Mitarbeiter als Koautoren auftauchen, ob die Autoren alphabetisch genannt werden, all das ist nicht wirklich konventionalisiert. Verschärfend kommt hinzu, daß Artikel, die in Mehrfachautorenschaft verfaßt wurden, in der Regel höhere Zitationsraten erreichen. Entscheidet man sich für das Prinzip „one paper is one paper“ und rechnet den einzelnen Autoren nur Bruchteile – entsprechend der Gesamtzahl der Autoren einer Publikation – zu, dann schlagen jedoch insbesondere jene die Sichtbarkeit erhöhenden und forschungspolitisch erwünschten Kooperationsbeziehungen negativ zu Buch.

Hat man diese Probleme gelöst bzw. sich für ein Verfahren entschieden, verfügt man über eine Art Aktivitätsindikator (der allerdings unter Umständen auch die impliziten oder expliziten Qualitätsbeurteilungen der Datenbankbetreiber reflektiert oder durch Gewichtungen implizite Qualitätszuweisungen enthält); wohlgemerkt einen disziplinspezifischen Indikator, denn die Länge von fachwissenschaftlichen Beiträgen und die Schreib- und Zitierhäufigkeit variieren erheblich zwischen den Forschungsgebieten.

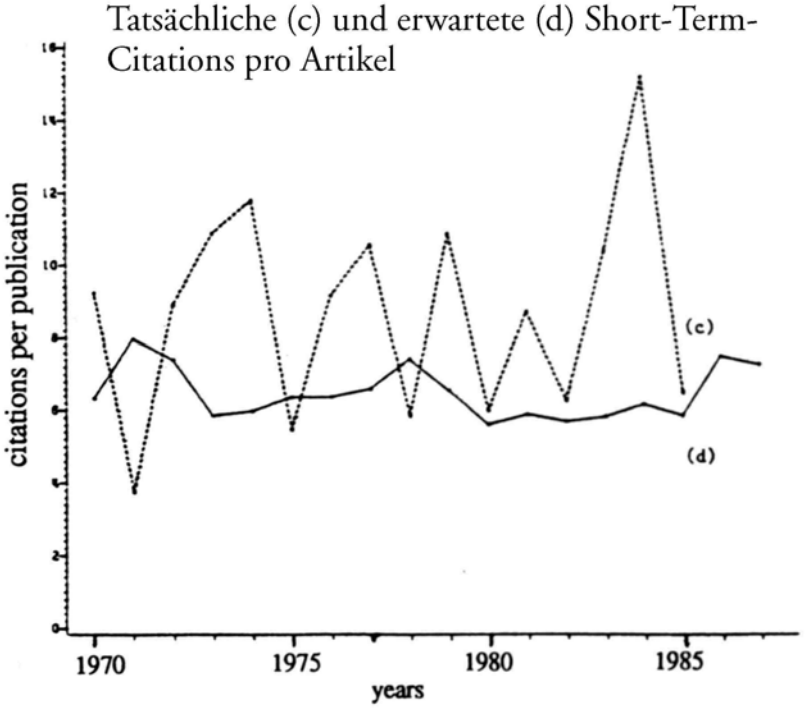
Will man von hier zu einer Qualitätsgewichtung kommen, benötigt man eine Einschätzung der wissenschaftlichen Bedeutsamkeit. Üblich ist eine solche Einschätzung einzelner Artikel im Rahmen der Manuskriptbegutachtung. Eine derartige Bewertung durch Peers ist jedoch für die enormen Mengen von Publikationen, die im Rahmen bibliometrischer Analysen verarbeitet werden müssen, nicht möglich. Das bekannteste Verfahren, dem einzelnen Artikel ein Gewicht – unabhängig von neuerlichen Bewertungen – zuzuweisen, ist die Ermittlung erhaltener Zitationen. Zitationsanalysen sind (wie übrigens alle Qualitätsgewichtungen) umstritten.

Zunächst einmal lassen sich nicht alle Zitate auffinden. Nur Referenzen aus Zeitschriften, die in einer der Datenbanken des ISI geführt werden, tauchen auch auf. Zweitens wird immer wieder die Frage gestellt, wofür ein Zitat eigentlich steht. Idealiter dokumentiert ein Zitat den Einfluß einer bestehenden Arbeit auf einen neuen Forschungszusammenhang. Tatsächlich reflektieren Zitationen natürlich neben Einfluß allerlei anderes. Zum Teil figurieren sie als sogenanntes Standard Symbol und bezeichnen einfach ein komplexes Verfahren oder eine Theorielinie, sodann

wirkt das Matthäus-Prinzip, das heißt, bekannte Autoren werden tendenziell häufiger zitiert, auch haben bestimmte Textgattungen (Überblicksartikel, Methodendarstellungen) häufig bessere Zitierchancen. Schließlich können auch innerhalb einer Disziplin sehr unterschiedliche Subkulturen hinsichtlich der Zitations- und Publikationsfrequenzen bestehen. Auf der anderen Seite werden in der Literatur allerdings immer wieder hohe Übereinstimmungen zwischen Peer-Urteilen, Auszeichnungen und anderen Forschungsindikatoren und Zitationsanalysen mitgeteilt. Beide Messungen gehen nicht ineinander auf, weisen aber deutliche Korrelationsbeziehungen auf. Insgesamt kann man sich wohl am ehesten darauf einigen, daß Zitationen die Sichtbarkeit von Publikationen abbilden. Zitationen liefern keine Qualitätsbewertung im methodologischen Sinne, sondern Informationen über Wahrnehmungen anderer Wissenschaftler.

Problematisch ist es weiterhin, zu entscheiden, wie tief die Retrospektive angelegt werden soll. Publikationen haben fachspezifisch sehr unterschiedliche Halbwertzeiten. Eine wichtige Publikation in der Mathematik wird häufig über Jahrzehnte zitiert, während gerade in neuen, sehr dynamischen Forschungsgebieten die Literatur sehr schnell veraltet und dann nicht mehr zitiert wird. Ein sogenanntes Zitationsfenster, das angibt, wie lange nach Erscheinen die Zitate registriert werden, muß also so definiert werden, daß einerseits auch einigermaßen aktuelle Publikationen erfaßt werden können, andererseits die Zitationsgewohnheiten der Disziplin berücksichtigt werden. Schließlich sind gegebenenfalls Selbstzitationen zu definieren und aus der Analyse zu entfernen.

Abbildung 4:



Quelle: van Raan (1994)

Beim interdisziplinären Vergleich ist weiterhin zu beachten, daß die Zitierhäufigkeiten nach Forschungsgebieten sehr unterschiedlich ausfallen. Damit ist die Frage berührt, was die Meßlatte für die Beurteilung der Indikatorwerte ist, oder anders formuliert, wie sich derartige Indikatoren normalisieren lassen. Häufig wird es notwendig sein, eine Bezugsgröße zu definieren.

Das kann z. B. das durchschnittliche Zitationsaufkommen der jeweiligen Zeitschrift sein oder das weltweite Zitationsaufkommen der Subdisziplin (wenn sie denn klar zu begrenzen ist). Es ist aber auch möglich, erhaltene Zitate anhand der vergebenen Referenzen zu normalisieren.

Die obige Abbildung zeigt z. B. einen Vergleich zwischen den (externen) Zitaten, die Publikationen aus einem Physik-Department der Universität Leiden erhielten (tatsächliche Zitationen), im Vergleich zu den Zitaten, die die übrigen Artikel in derselben Zeitschrift im Durchschnitt weltweit erhielten (erwartete Zitationen). Damit läßt sich zumindest grob identifizieren, ob eine Publikation durchschnittliche bzw. über- oder unterdurchschnittliche Resonanz in der Scientific community erhielt.

Ein anderer Weg zur Qualitätseinschätzung benutzt pauschale Beurteilungen von Artikeln derart, daß von Wissenschaftlern die Wichtigkeit oder das Qualitätsniveau einzelner Zeitschriften erfragt wird und auf dieser Grundlage dann ein Zeitschriftset ausgewählt wird oder den einzelnen Artikeln Gewichtungen zugewiesen werden. Solche pauschalen Gewichtungen lassen sich auch mit Hilfe bibliometrisch gewonnener Gewichtungsfaktoren bewerkstelligen. Grundsätzlich werden dazu die vergebenen Referenzen und erhaltenen Zitate einer ganzen Zeitschrift verrechnet. Die daraus entwickelten Gewichtungsfaktoren unterscheiden sich nur in den Details ihrer Konstruktion (Impact Factor, Influence Weight). Das Problem solcher Gewichtungen liegt vor allen Dingen in der Varianz innerhalb der Zeitschriften: Auch sehr angesehene Zeitschriften weisen erhebliche Unterschiede in der Zitationshäufigkeit der einzelnen Beiträge auf. Dem einzelnen Beitrag wird man daher mit einer Zeitschriftengewichtung in der Regel nicht gerecht. Bei größeren Publikationsmengen (über 100) nähern sich jedoch individuelle Zitationsgewichte und Zeitschriftengewichte an, wie auch verschiedene Gewichtungsverfahren dann tendenziell konvergieren.

Die Auswirkungen von Entscheidungen für oder gegen bestimmte Gewichtungs- und Auswahlverfahren werden besonders deutlich sichtbar, wenn aus den ermittelten Daten Effizienzmaße konstruiert werden. Die folgende Tabelle zeigt, daß, je nachdem ob man nur Publikationen zählt, Gewichtungen anhand von erhaltenen Zitaten vornimmt oder berücksichtigt, wie viele der Autoren tatsächlich aus dem analysierten Fachbereich stammen, mal die Konstanzer Physiker erfolgreicher sind, mal die Kölner, die sehr viel mehr Artikel in Kooperation mit Wissenschaftlern anderer Universitäten verfaßt haben.

Tabelle 1:

Vergleich verschiedener Effizienzmaße Publikationen (1983-88) und Zitationen (in den ersten 3 Jahren nach Veröfflichung) aus dem Forschungsgebiet Physik der Universitäten Köln, Konstanz und Bremen							
Universität	I Zahl der Publikationen *	II Mit Zitaten gewichtete Publikationen **	III Mit Zitaten gewichtete Autorenanteile ***	IV Personal am Fachbereich ****	Relationen		
					I / IV	II / IV	III / IV
Köln	662	1.455	440,9	47,3	14,0	30,8	9,3
Konstanz	316	725	441,9	24,8	12,7	29,2	17,8
Bremen	82	143	82,1	22,7	3,6	6,3	3,6

- alle Publikationen der Datenbank Scisearch, die als institutionelle Adresse den Fachbereich Physik oder entsprechende Institute und Lehrstühle aufweisen.

** Gewichtungsfaktoren: 0-1 Zitat: 1; 2-5 Zitate: 2; 6-10 Zitate: 3; 11-15 Zitate: 4; 16-20 Zitate: 5; >20 Zitate: 6 .

*** Gewichtungsfaktor: Autoren aus dem Fachbereich / Gesamtzahl der Autoren.

**** Professoren wurden mit 1, Mittelbaustelle mit 1/2 und Qualifikationsstellen mit 1/3 gewichtet. Teilzeitstellen wurden als halbe Stellen gerechnet.

Quelle: Hornbostel (1997)

Drittmittel

Zu den wichtigsten und gebräuchlichsten Forschungsindikatoren, die nicht auf Publikationen zurückgreifen, gehören die Drittmitteleinwerbungen. Sie sind vor allen Dingen deshalb interessant, weil sie anders als etwa bibliometrische Indikatoren, die nur retrospektiv berichten, sehr zeitnahe oder sogar prospektive Informationen liefern. Bei diesem Indikator ist allerdings strittig, ob man ihn als Input- oder Output-Messung interpretieren kann. Für erstere Deutung spricht, daß die Mitteleinwerbungen noch nichts über den Ertrag eines Projekts aussagen. Für letztere spricht, daß erstens der Drittmittelbewilligung in der Regel ein aufwendiger Begutachtungsprozeß zugrunde liegt, zweitens mit dem Drittmittelantrag häufig bereits ein sehr voraussetzungsvolles Zwischenprodukt vorliegt. Ein Indiz dafür, daß die Begutachtung von Drittmittelanträgen durchaus prognostische Validität besitzt, ergibt sich, wenn man einmal den Publikationsoutput daraufhin untersucht, ob die aus drittmittelgeförderten Projekten hervorgegangenen Publikationen eine andere fachliche Resonanz erzeugen als die übrigen Publikationen.

Tabelle 2:

Publikationen (1983-87) und Zitate in den ersten drei Jahren nach Erscheinen für die physikalischen Fachbereiche der Universitäten TU Berlin, Köln und Konstanz							
Mit Drittmittelförderung				Ohne Drittmittelförderung			
Zahl der Artikel *	Mean Zitate **	Nicht Zitiert ***	Viel zitiert ****	Zahl der Artikel *	Mean Zitate **	Nicht Zitiert ***	Viel Zitiert ****
543	7,4	127 (23,4%)	42 (7,7%)	469	4,9	171 (36,5%)	14 (3,0%)

* : Zahl der in SCISEARCH nachgewiesenen Artikel (nach institutioneller Adresse).

** : Mittelwert (mean) der in den ersten 3 Jahren nach Erscheinen erhaltenen Zitate.

*** : Artikel, die in den ersten 3 Jahren nicht oder einmal zitiert wurden.

**** : Artikel, die in den ersten 3 Jahren mehr als 20 mal zitiert wurden

Quelle: Hornbostel (1997)

Im Durchschnitt erreichen drittmittelgeförderte Projekte höhere Zitationsraten, sie erreichen geringere Anteile von nicht zitierten und höhere Anteile von stark zitierten Arbeiten. Auch zwischen der Produktivität der Fachbereiche (gemessen in Publikationen) und den eingeworbenen Drittmitteln bestehen deutlich positive Korrelationen, unabhängig davon, ob man mit absoluten Werten rechnet oder mit Pro-Kopf-Angaben.

Tabelle 3:

Pearsons Corr. für Drittmittel und Publikationen (Bundesrepublik Deutschland,1983-88)			
Fach	Drittmittelbe-willigungssumme (DFG, DFG-Sonderforschungsber., BMFT, VW)	Publikationen des Fachbereichs in SCISEARCH 1983-1988 absolut	Publikationen des Fachbereichs in SCISEARCH 1983-1988 je Professor
Biologie (44 Fachbereiche)	Absolut	**.77**	
	je Professor		**.69**
Chemie (45 Fachbereiche)	Absolut	**.86**	
	je Professor		**.66**
Physik (46 Fachbereiche)	Absolut	**.70**	
	je Professor		**.67**

Anm.: Nur Lehr- und Forschungsbereiche mit durchschnittlich mehr als einem Professor 1984 und 1986.

Quelle: Hornbostel (1997)

Es spricht also viel dafür, Drittmitteleinwerbungen, insbesondere in Ergänzung zu Publikationsanalysen, als Forschungsindikator zu benutzen. Voraussetzung ist allerdings, daß Drittmittelforschung in den untersuchten Fachgebieten üblich ist und nicht ausschließlich im Rahmen subdisziplinärer Spezialisierung (insbes. empirische Forschung) betrieben wird. In den naturwissenschaftlichen und technischen Disziplinen wird durchschnittlich von jedem Professor etwa ein Drittmittelprojekt im Semester betreut, in den Wirtschafts- und Sozialwissenschaften ist der Umfang der Drittmittelforschung immer noch nennenswert, wenngleich sehr viel niedriger, in den Rechtswissenschaften wird Drittmittelforschung nur in wenigen Spezialgebieten betrieben.

Dieser Sachverhalt zeigt sich auch, wenn man Professoren nach ihren Antragsaktivitäten fragt: Wie die folgende Tabelle zeigt, haben in der jüngsten Umfrage (Vollerhebung mit ca. 43% Rücklauf) des CHE (Centrum für Hochschulentwicklung) nur ca. 14% der Physikprofessoren angegeben, in den Jahren 1995 bis 1997 kein Drittmittelprojekt beantragt zu haben, während der Anteil unter den Juristen sich auf fast 64% beläuft. Umgekehrt haben in der Physik und der Informatik 12 bis 13% der Professoren – nach eigenen Angaben – mehr als zehn Förderungsanträge gestellt. Wegen der geringen Antragsintensität und der starken Fokussierung auf Spezialgebiete macht es keinen Sinn, die Drittmitteleinwerbungen in den Rechtswissenschaften als Indikator für Forschungsleistungen zu benutzen.

Tabelle 4:

Professoren nach Zahl der Drittmittelanträge 1995-1997				
Zahl der Drittmittel-anträge	Physik	Informatik	Mathematik	Jura
0	14,4 %	18,0 %	38,4 %	63,6 %
1-5	48,6 %	45,4 %	49,5 %	30,8 %
6-10	23,7%	24,8 %	9,4 %	4,2 %
> 10	13,2 %	11,8 %	2,7 %	1,4 %
Gesamt	100 %	100 %	100 %	100 %

Quelle: CHE (1999), Professorenbefragung

Weiterhin sollten die eingeworbenen Drittmittel danach differenziert werden, ob sie einem Begutachtungsprozeß unterlagen oder nicht. Als Qualitätsindikator eignen sich nur Mittel, die aufgrund einer Begutachtung und Empfehlung von Peers vergeben wurden. Das heißt auch, daß nicht alle Zuwendungen (z. B. Druckbeihilfen, Stipendien etc.), sondern nur Mittel, die in unmittelbare Forschungsaktivitäten fließen, zu einem Indikator verarbeitet werden sollten. Schließlich ist darauf zu achten, daß Drittmittel, die verschiedenen Universitäten zugute kommen – wie z. B. bei Sonderforschungsbereichen – auch entsprechend zugerechnet werden bzw. Mittel, die lediglich verwaltet oder weitergereicht werden, auch nur bei den Empfängerinstitutionen auftauchen.

Patentindikatoren

Grundsätzlich werden anwendungsnahe Forschungsgebiete durch bibliometrische Indikatoren nur schlecht abgebildet. Patentindikatoren können dies in gewissem Maße kompensieren, nicht nur in technischen Forschungsgebieten, sondern zunehmend auch in speziellen Bereichen der Biowissenschaften.

Da Patentanmeldungen jedoch in erster Linie auf die Sicherung wirtschaftlicher Verwertungsansprüche zielen und der Patentierungsprozeß selbst aufwendig und kostspielig ist, bedarf es jeweils einer sehr genauen Prüfung der Aussagekraft eines solchen Indikators für das untersuchte Gebiet. Immerhin weisen die ingenieurwissenschaftlichen Forschungsgebiete (insbesondere Maschinenbau/Verfahrenstechnik) und die verschiedenen Spezialgebiete der Chemie (mit Einschränkungen auch Medizin, Elektrotechnik, Physik und Biologie) in einem Umfang Patentanmeldungen aus dem Hochschulkontext auf, der es rechtfertigt, diese Patentierungsaktivitäten zu einem Indikator zu verarbeiten.

Erhebliche Schwierigkeit bereitet allerdings die Identifikation der institutionellen Zugehörigkeit der Erfinder, hier bestehen (aufgrund des sogenannten „Hochschullehrerprivilegs“) keine klaren, einheitlichen Konventionen oder gar Verpflichtungen hinsichtlich der Beteiligung der Universität am Anmeldeverfahren, so daß es kaum möglich ist, zuverlässige Angaben über die Hochschul- und Fachbereichszugehörigkeit der Erfinder aus den Patendatenbanken zu erhalten. Obwohl die Hochschulen und die außeruniversitären Forschungseinrichtungen in letzter Zeit die Patentanmeldungen ihres Personals immer stärker fördern und auch häufiger als Anmelder auftreten, ist eine Identifikation der Erfinder bisher nur über die Namen möglich. In der Folge entstehen alle jene aus Publikationsanalysen bekannten Probleme der Zuordnung von Personen zu Institutionen (Homonyme, wechselnde Zugehörigkeiten etc.). Schließlich bleibt als Problem, daß die Patentanmeldung keineswegs mit einem nachhaltigen Verwertungsinteresse bzw. einer nachhaltigen Verwertungschance gleichzusetzen ist.

Unter analytischen Gesichtspunkten bieten Patentindikatoren jedoch interessante Verknüpfungen mit bibliometrischen Analysen, da in Patentschriften auch wissenschaftliche Literatur zitiert wird. Allerdings bewegen wir uns damit schon auf dem Feld der Strukturindikatoren, mit denen Transferprozesse zwischen Forschung und Anwendung abgebildet werden können.

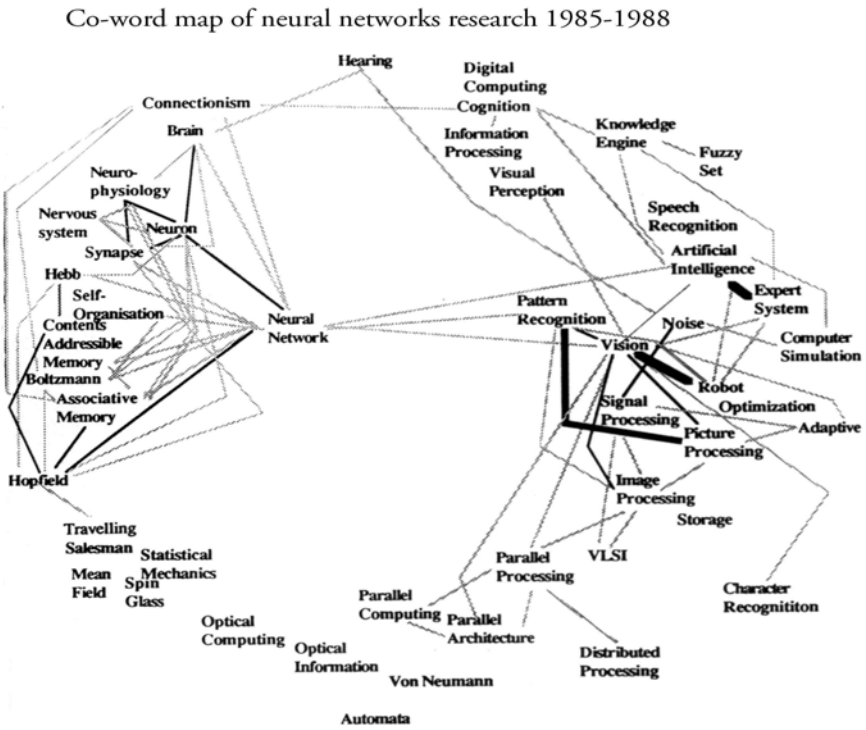
Strukturindikatoren

Strukturindikatoren zielen weniger auf eine Outputmessung als auf eine Strukturbeschreibung. Die Grundlage solcher Indikatoren sind entweder Publikationsanalysen, die beispielsweise erlauben, Kooperationsbeziehungen von Autoren oder Institutionen darzustellen, Zitationsanalysen , die die Identifikation von aktuellen „Forschungsfronten“ ermöglichen, oder sog. Co-Word Analysen, die aus Titeln, Abstracts oder ganzen Texten Clusterungen von Texten nach inhaltlichen Gemeinsamkeiten vornehmen. Möglich sind ebenso Analysen, die Interdisziplinarität oder Internationalität zum Gegenstand haben oder Vernetzungsstrukturen beschreiben.

Häufig entstehen dabei nicht unmittelbar Indikatoren, sondern Kartierungen („Mapping“) von Forschungsgebieten. Allerdings lassen sich solche Abbildungen von Strukturen in der Forschungslandschaft auch unter Leistungsgesichtspunkten interpretieren. Die Teilhabe an sich schnell entwickelnden, aktuellen Forschungsschwerpunkten kann z. B. als besondere Leistung gewertet werden.

Die folgende Abbildung zeigt die thematische Vernetzung eines Forschungs-schwerpunkts. Die Karte beruht auf dem gemeinsamen Auftreten von Schlüssel-begriffen im Abstract, Titel oder Text eines Artikels und bezeichnet die „Nähe“ der Forschungscluster zueinander. Innerhalb dieser Cluster können nun wiederum sehr aktive Forschergruppen identifiziert werden.

Abbildung 5:



Quelle: van Raan (1994)

Andere Indikatoren

Sichtet man die Versuche, Forschungsleistungen meßbar zu machen, taucht eine Reihe weiterer Indikatoren auf, die hier nicht im einzelnen diskutiert werden können. Am gebräuchlichsten sind Angaben zur Förderung des wissenschaftlichen Nachwuchses. Insbesondere die Anzahl von Promotionen und Habilitationen wird gern benutzt. Dieser Indikator ist aber mit Vorsicht zu behandeln. Die Anforderungen an ein Promotionsverfahren sind zunächst einmal disziplinspezifisch. Sie liegen in jedem Fall aber auf der Grenze zwischen Lehr- und Forschungsaktivität, so daß es sich nicht um einen reinen Forschungsindikator handelt. Zudem bleibt ausgesprochen unklar, ob die Zuordnung der Promotionen zu Institutionen eher akzidentell ist oder tatsächlich institutionelle Leistungen beschreibt.

Auch Einladungen, Vorträge, Stipendien, Herausgeberschaften, Gutachtertätigkeiten u. ä. werden gelegentlich ausgezählt und als Indikator benutzt. Sicherlich hat all dies mit Reputationszuweisungen im Wissenschaftssystem zu tun; um verläßliche Indikatoren konkreter Forschungsleistungen handelt es sich jedoch meist nicht. Hier muß sehr genau geprüft werden, ob derartige Indikatoren in erster Linie

Vernetzungintensitäten abbilden oder zu gleicher Zeit Forschungsbeiträge indizieren. Zumindest sollten solche Informationen durch andere Indikatoren substantiiert werden.

Resümee

Bei Forschungsindikatoren handelt es sich um Partialindikatoren, und zwar in einem doppelten Sinne: Sie bezeichnen die Partialität des Meßkonzepts im Hinblick auf "wissenschaftlichen Fortschritt" und die partielle Indizierung des Konzepts durch Indikatoren. Schließlich ist jeder Indikator mit methodisch-technischen Problemen behaftet, die sowohl Entscheidungen vom Konstrukteur des Indikators verlangen (was ist eine geeignete Datenbasis, was sind sinnvolle Vergleichsmaßstäbe, welche Relationen sind angebracht, um reine Größeneffekte auszuschalten?) als auch erhebliche Kontrollen, Korrekturen und Modifikationen notwendig machen.

Angesichts der vielfältigen Einschränkungen drängt sich die Frage auf, ob man denn überhaupt guten Gewissens Forschungsindikatoren benutzen kann. Um Mißverständnissen vorzubeugen: Die Tatsache, daß wir inzwischen relativ viel über die Stärken und Schwächen solcher Indikatoren wissen, spricht nicht gegen eine Verwendung, sondern nur gegen einen naiven Umgang mit derartigen Indikatoren. Der Aufforderung einer „Rechenschaftslegung“ kann sich eine öffentlich finanzierte Forschung nicht entziehen. Und sie tut gut daran, eigenständig ein Bündel von Indikatoren zur Verfügung zu stellen, das geeignet ist, die Schwächen der einzelnen Indikatoren zu kompensieren und zugleich als Konvention über relevante Dimensionen der Forschungstätigkeit akzeptiert zu werden.

In ihren Eigenheiten und Schwächen unterscheiden sich Forschungsindikatoren nicht von anderen ökonomischen oder sozialen Indikatoren. Sie alle ähneln einer Landkarte, die immer Kompromisse eingehen muß, wenn die dreidimensionale Realität auf die zweidimensionale Karte übertragen wird. Die Karte gibt nun zweifellos etwas von der Realität des Gebiets wieder, selbst dann, wenn einige Nutzer ihren Orientierungsbedarf nicht befriedigt sehen. Mehr noch, das Lesen der Karte verlangt besondere Qualifikationen, und zur Rückübersetzung der Karte in Orientierungswissen sind häufig zusätzliche Informationen notwendig, die die Karte nicht enthält (z. B. reichen die Entfernungsangaben nicht aus, den Zeitaufwand abzuschätzen, um von A nach B zu kommen).

Ganz ähnlich verhält es sich mit der Nutzung von Leistungsindikatoren. Dem üblichen Beurteilungsprozeß kommt das kontextsensitive, wenig formalisierte Urteil von Expertengruppen am nächsten. Der Vorteil liegt darin, daß Experten situationsadäquate Relevanzspezifizierungen vornehmen. Der Nachteil dieses Verfahrens besteht – neben den Grenzen der Belastbarkeit der Peers – in der nicht hintergehbaren Subjektivität und damit der Möglichkeit eines unangemessenen Urteils. Auf der anderen Seite stehen "objektive" Messungen, die weitgehend

situationsunabhängig nach einem festen Verfahren Informationen und dann auch Bewertungen produzieren. Gerade den genauesten Messungen, etwa der Auszählung von Publikationen, fehlt aber eine klare Spezifikation der Relevanz der so gewonnenen Informationen. Objektivität im Sinne genauer und wiederholbarer Messungen ist ähnlich wie im Falle der Landkarte ein notwendiges, aber keinesfalls hinreichendes Gütekriterium. Insoweit ist die Verwendung von Indikatoren einerseits riskant, andererseits unumgänglich, wenn sehr komplexe Sachverhalte vergleichend beurteilt werden sollen.

Forschungsindikatoren sind ursprünglich als analytische Instrumente, nicht als Leistungsindikatoren entwickelt worden. Sie sind daher in sehr unterschiedlichen Kontexten nutzbar. Sie lassen sich sowohl zur Vertiefung und Erweiterung des Reflexionswissens einsetzen (Wissenschaftsforschung) wie auch als Selbstevaluation, d. h. als ein freiwilliges "Monitoring" von Institutionen oder Forschergruppen. Schließlich stellen Wissenschaftsindikatoren ein wissenschaftspolitisches Kontroll-, Informations- und Steuerungspotential dar, werden also bei der Aushandlung akzeptanzfähiger Elemente eines Berichtssystems wichtig. Ein solides Indikatorenprogramm sollte so eingerichtet sein, daß es an die Entwicklungen der Wissenschaftsforschung rückgebunden bleibt und Informationen bereitstellt, die sowohl auf der untersten Erhebungsebene steuerungsrelevant sind als auch einen Leistungsvergleich mit anderen Institutionen ermöglichen.

Ein derartiges Berichtssystem ist keine Einschränkung, sondern wohl eher Garant wissenschaftlicher Autonomie. Gefährdet wird diese Autonomie erst, wenn es der Wissenschaft nicht gelingt, überzeugend über die eigene Leistungsfähigkeit zu informieren. Dann greifen, wie die hochschulpolitischen Entscheidungen der letzten Jahrzehnte lehren, andere bürokratische Urteilskriterien.

Literaturverzeichnis

CHE Centrum für Hochschulentwicklung (1999), Studienführer, Gütersloh

Hornbostel, S. (1997), Wissenschaftsindikatoren, Opladen

van Raan, A. F. J. (1994), Assessment of Research Performance with Bibliometric Methods, in: Best, H. et. al. (Hg.), Informations- und Wissensverarbeitung in den Sozialwissenschaften, Opladen, S. 499-524

Dagmar Simon

Wer evaluiert zu welchem Zweck was?

Anmerkungen zu Zielen und Verfahren der Selbstevaluation in außeruniversitären Forschungseinrichtungen

Wenn von Selbstevaluationen, Selbstkontrolle und Selbstbeobachtungssystemen außeruniversitärer Forschungseinrichtungen die Rede ist, sollte auch ein Wort zu dem „Pendant“ – externe Forschungsevaluierungen – verloren werden, insbesondere vor dem Hintergrund, daß in der Bundesrepublik weder auf eine besonders elaborierte wissenschaftspolitische Debatte über Aufgaben und Zielen von externen und internen Evaluationen in diesem Sektor noch auf eine auf Praxis zurückgeblickt werden kann, die auf Kontinuität, Routinisierung und Weiterentwicklung von Modellen und Verfahren beruht. Gerade die Evaluation der Einrichtungen der Blauen Liste durch den Wissenschaftsrat zeigt, daß diese Bewertungsprozesse für viele der betroffenen Institute immer noch eine "Ausnahmesituation" im Forschungsalltag und weniger eine Routine im Kontext ihrer Forschungspraxis darstellen. Zudem findet dieses Thema bisher noch wenig Resonanz in der Wissenschafts- und Evaluationsforschung oder gar in einer sozialwissenschaftlichen Begleitforschung. Somit ist es auch erklärbar, daß bei diesem Thema auf Selbstevaluationsmodelle im *Hochschulbereich* und auf internationale Erfahrungen rekurriert werden muß (vgl. die Beiträge von Jürgen Lüthje und Adrian Verkleij in diesem Paper).

Die Diskussion über Ziele und Verfahren von Forschungsbewertungen hat in der jüngsten Zeit neue Anstöße erhalten – zunächst durch die Evaluation der Institute und Forschungsgruppen der Akademie der Wissenschaften der DDR und danach durch die Evaluierung von über 80 Forschungs- und Serviceeinrichtungen der Blauen Liste durch den Wissenschaftsrat. Vor dem Hintergrund der „Empfehlungen zur Neuordnung der Blauen Liste“ wurde diese Evaluierung unter einem *einheitlichen* wissenschaftspolitischen Auftrag sowie einheitlicher Bewertungskriterien zur Erfassung der Institutsleistungen durchgeführt.

Aus der Perspektive des Projekts „Institutionelle Selbstbeobachtung als Steuerungsinstrument für außeruniversitäre Forschungseinrichtungen?“ (vgl. Einleitung S. 7 ff.) soll in diesem Beitrag zunächst kurz auf einige aktuelle Debatten an den Hochschulen über Lehr- und (am Rande) Forschungsevaluationen eingegangen werden, die für unsere Fragestellungen relevant erscheinen. Dann werden die Sichtweisen der untersuchten Blaue Liste-Institute kommentiert, vor allem hinsichtlich der Erwartungen an externe Bewertungen und interne Selbstbeobachtungsprozesse sowie die damit verbundenen Aufgaben und Zielvorstellungen. Sie werden zum einen vor dem Hintergrund der Evaluation dieser Institute durch den Wissenschaftsrat und zum anderen ihrer unterschiedlichen und unterschiedlich ausgeprägten Praktiken zur Qualitätssicherung und Förderung diskutiert. In einem zweiten Schritt werden Elemente eines für die Evaluation der

Institute der Wissenschaftsgemeinschaft G.W. Leibniz (WGL) zu konzipierenden zukünftigen Modells skizziert.

In der Diskussion über Evaluationen an den Hochschulen fällt auf, daß erstens die Frage nach ihren Zielen noch nicht (befriedigend) beantwortet ist: "Mehr Transparenz durch validere Information", "Rechenschaftslegung", "Qualitätssicherung und -verbesserung" sind die Favoriten aus der Sicht der Hochschulakteure, weniger: "Vergleich", "Ranking" und "Steuerung der Ressourcen und Finanzen" (vgl. Hochschulrektorenkonferenz 1998).

Dabei kristallisieren sich Transparenz und Rechenschaftslegung ("accountability") der Hochschulen einerseits und Verbesserungsmöglichkeiten ("improvement") von Lehre und Studium andererseits als die Debatte beherrschenden Schlagworte heraus. Angemerkt sei hier, daß die wissenschaftspolitischen Akteure Rechenschaftslegung und Transparenz auch von *außeruniversitären* Forschungseinrichtungen verlangen. Das Verhältnis der beiden Ziele zueinander wird unterschiedlich charakterisiert: In der internationalen Diskussion werden "accountability" und "improvement" als entgegengesetzte Pole – Scylla und Charybdis – (Vroeijenstijn 1995, S. 339) gesehen, von anderer Seite (Barz et al. 1997) wird den meisten nationalen Evaluierungssystemen eine Verbindung von öffentlicher Rechenschaftslegung und einer an den Zielen und Konzepten der Hochschule orientierten Qualitätsentwicklung, wenn auch unterschiedlicher Ausprägung, zugeschrieben.

Zweitens ist auch die Deutung des Qualitätsbegriffs je nach Akteuren und spezifischen Interessenkonstellationen unterschiedlich, vor allem hinsichtlich der mit Evaluationen verbundenen Zielvorstellungen und ihrer Konsequenzen. Harvey und Green fassen dies treffend wie folgt zusammen: "Quality as exceptional, quality as perfection or consistency, quality as fitness for purpose, quality as value for money, quality as transformation. This grouping spans the universe from traditional notions of academic quality as excellence ('exceptional') to the recent insights of 'zero defects' ('perfection'), mission orientation and consumer orientation ('fitness for purpose'), and finally, in the transformational notion of quality, to the question of what higher education is about" (vgl. Westerheijden et al. 1994, S. 16/17).

Drittens wird in neueren Veröffentlichungen, insbesondere in denjenigen zum Projekt Qualitätssicherung der Hochschulrektorenkonferenz (1998 und 1999), der Zusammenhang von Evaluierungsverfahren und Struktur-, Organisations- und Managementfragen hervorgehoben, d.h. Evaluierungen werden in den Kontext von Organisations- und Entwicklungsfragen von Universitäten eingeordnet. Es geht um die Identifizierung von Organisationszielen und, davon abgeleitet, um geeignete Organisationsstrukturen, um Anpassungsprozesse der Aufbau- und Ablauforganisation, um Kundenorientierung (bspw. ein Verständnis von Studenten als "Kunden"). Dabei erstrecken sich die Ansätze von der Evaluation der Lehre als Organisationsberatung durch Hochschulforscher an der Bielefelder Universität (vgl. Webler 1998) bis hin zu Versuchen, Total Quality Management (TQM)-Verfahren

und das European Model for Business Excellence an Hochschulen zu implementieren.

In diesem Zusammenhang werden bestimmte Evaluationsmodelle, nämlich prozeßorientierte mit einer Schwerpunktsetzung auf Optimierungsfunktionen, favorisiert und gleichzeitig die Bedeutung von Selbstbewertungsmodellen als Voraussetzung für erfolgreiche Qualitätssicherung und -förderung herausgestellt. „Professionelle Selbstkontrolle muß langfristig den Kern einer Strategie der Entwicklung von universitärer Qualität ausmachen, weil 'Entwicklung' immer auf die Mitarbeit der Institutionsmitglieder, auf ihre Motivationen und die differenzierteren qualitativen Informationen, die die Innensicht erlaubt, angewiesen ist" (Altrichter et al., 1997, S. 13).

In unserer Untersuchung schenken wir den Sichtweisen, spezifischen Ausgangssituationen und Problemlagen der WGL-Institute besondere Aufmerksamkeit. Angesichts der Heterogenität der Institute – hier kommen unterschiedliche Organisationstypen, Funktionen, Wissensstrukturen, Forschungstechniken und Bezugsgruppen ins Spiel – müssen, nach Einschätzung der Institutsakteure, Evaluationen die Besonderheiten der Institute berücksichtigen und überprüfen. Dies impliziert auch eine institutsadäquate Gestaltung des Verfahrens; beispielsweise die Einbeziehung von Nutzern der Forschung in den Evaluiererkreis. Erwartet wird ein hoher Anteil an beratenden Funktionen bei externen Evaluationen und eine stärkere Rückkopplung zwischen Evaluatoren und Evaluierten. Forschungsbewertungen sollen eine kritische Reflexion der Standards bewirken, eine Lokalisierung von Schwachstellen ermöglichen und Entwicklungsperspektiven eröffnen. Außerdem wird die Notwendigkeit gesehen, Indikatoren für Fehlentwicklungen zu ermitteln. Viele Institutsvertreter messen dem Forschungsmanagement eine hohe Bedeutung bei; dementsprechend sollte beispielsweise der Personalpolitik – insbesondere der Mitarbeiterführung – in Evaluationen Beachtung geschenkt werden.

Institutionellen Selbstbeobachtungsverfahren wird einerseits eine Relevanz als Voraussetzung für eine interne Qualitätssicherung, die Vorbereitung von externen Evaluierungen und als integraler Bestandteil des Forschungsprozesses beigemessen. Dennoch besteht eine gewisse Skepsis der Institutsakteure gegenüber dem Nutzen von internen Evaluierungen, da diese Verfahren in der Regel nicht auf Anreiz- und Sanktionssystemen basieren und die Gefahr einer Wirkungslosigkeit dieses Instruments gesehen wird. Eine gewisse Unentschiedenheit und Unsicherheit hängt hier möglicherweise auch mit dem Verweis auf andere Elemente der Qualitätssicherung in den Instituten wie interne Refereessysteme, Publikationskommissionen, Vorabpräsentationen von Vorträgen oder Beratungen bei der Drittmittelantragsstellung zusammen. Hier ergibt sich unseres Erachtens die Chance und auch Notwendigkeit, diese Verfahren und Strukturen selber als Gegenstand von institutsinternen Evaluierungen hinsichtlich ihrer Funktionalität und Reichweite zu begreifen.

Die Bedeutung einer kontinuierlichen Überprüfung der Leistungsfähigkeit und der strukturellen Rahmenbedingungen kommt unter anderem in den Vorstellungen über die Aufgaben der wissenschaftlichen Institutsbeiräte zum Ausdruck: Sie sind primär als Beratungsinstanz für strategische Orientierungen, bei der Entwicklung von Zukunftsperspektiven und bei der Identifizierung von Forschungstrends gefragt. Der Beirat soll durchaus "Finger in die offenen Wunden legen", d.h. sich kritisch mit den Institutsleistungen auseinandersetzen.

Die ingesamt mit Evaluationen verbundenen Erwartungen seitens der Institutsleitung und auch der wissenschaftlichen Mitarbeiter/innen sind hoch: Eine systematische Differenzierung zwischen der Aufgabenstellung, den Akteuren und Adressaten interner Verfahren einerseits und externer Verfahren andererseits bzw. die gegenseitige Bezugnahme ist jedoch erst in Ansätzen entwickelt. Auch mit Blick auf die Entwicklung eines institutsangemessenen Indikatorensystems ist die Frage nach unterschiedlichen externen und internen Bedarfen noch zu beantworten.

Die Wissenschaftsratsevaluationen werden – wie schon oben angedeutet – überwiegend als "Ausnahmezustand" hinsichtlich der Bewältigung von Institutsaufgaben begriffen; eine "Evaluationskultur" scheint noch nicht in Sicht. Dabei ist die Vorbereitung auf die Evaluierung als produktiv empfunden worden, da sich die wissenschaftlichen Mitarbeiterinnen und Mitarbeiter in einer *systematischen* Weise mit dem Aufgaben- und Leistungsprofil des Instituts auseinandersetzen und "über den Tellerrand" des eigenen Forschungsbereichs blicken mußten. So bewirkte der Evaluierungsprozeß insgesamt eine größere Transparenz der verschiedenen Aufgaben- und Leistungsbereiche; Defizite wurden institutsweit sichtbar und Diskussionsgegenstand sowie ein Problembewußtsein insbesondere für eine verstärkte Kooperation zwischen den Forschungseinheiten wurde gefördert. Die Problematisierung des hohen Zeitaufwands für die Vorbereitung weist auf Defizite institutsinterner Kommunikations-, Selbstverständigungs- und Selbstkontrollprozesse im Sinne eines routinisierten Verfahrens hin. Auch die Auswertung der Erfahrungen mit den Evaluationsverfahren – sofern diese überhaupt stattgefunden hat – wurde nur punktuell auf Ziele, Methoden und Verfahren interner Selbstevaluationssprozesse und damit auf Fragen der Selbststeuerung und Organisationsentwicklung von Forschungsinstituten bezogen.

Die Institutsakteure empfanden mehrheitlich die Möglichkeiten zur Darstellung der Forschungsarbeiten und zur Diskussion mit den Gutachtern im Rahmen der Begehung als nicht ausreichend (der produktive wissenschaftliche Austausch mit den Evaluatoren sei zu kurz gekommen); zum Teil wurde die universitär geprägte Sichtweise der Gutachter – gerade bei einer stark anwendungsorientierten und häufig politikbezogenen Forschung mit einem erheblichen Anteil an Service- und Beratungsleistungen – als problematisch gewertet. Aus unserer Sicht ist bei der Konzeption zukünftiger Evaluationsverfahren für die WGL darüber nachzudenken, inwieweit Evaluationsverfahren mehr im Sinne eines interaktiven Prozesses gestaltet

werden können und ob Nutzer der Forschungsergebnisse in den Kreis der Evaluierer aufgenommen werden sollten.

Da die Wissenschaftsratsevaluationen nicht nur auf eine Bewertung der Forschungs- und Serviceleistungen der Blaue Liste-Institute abzielen, sondern auch Empfehlungen zur wissenschaftlichen Ausrichtung des Instituts und seiner institutionellen Voraussetzungen und Rahmenbedingungen ausgesprochen werden, sollten Verfahrenselemente – wie die Institutsbegehung – hinsichtlich der Förderung von Lernprozessen in den Instituten überdacht werden. Die beim Verfahren des Wissenschaftsrats einzukalkulierenden zeitlichen und personellen Restriktionen bei der Evaluierung von über 80 Blaue Liste-Institute spielen möglicherweise bei einem von den WGL-Gremien zu konzipierenden und durchzuführenden Verfahren eine nicht so bedeutende Rolle. Insbesondere sollte den Instituten nach Vorlage des Evaluationsberichts eine Feed-back-Runde mit den Evaluatoren ermöglicht werden, die im Sinne eines interaktiven Prozesses gestaltet werden könnte.

Abschließend werden hier einige Elemente eines zukünftigen Evaluationsmodells vorgestellt, die vor allem für die Blaue Liste-Institute angemessen erscheinen:

1. Mit wissenschaftsgeleiteten Evaluationen im Forschungsbereich sollte sowohl eine Bewertung der Leistungen und ihrer institutionellen Rahmenbedingungen als die *Förderung der Qualität der Forschung und damit auch ihrer Voraussetzungen* intendiert werden – insbesondere angesichts neuer Anforderungen an das Wissenschafts- und Forschungssystem. Wichtiges Desiderat stellt hierbei die Entwicklung eines qualitätsfördernden Evaluationsmodells für außeruniversitäre Forschung zur Sicherung und Gewährleistung der Leistungs- und Lernfähigkeit von Forschungsinstituten dar.

2. Angesichts der Vielfältigkeit und Unterschiedlichkeit der Aufgaben, Ziele, Organisationstypen, Wissensstrukturen und Bezugsgruppen den WGL-Institute müssen Evaluationen ihre Spezifika erfassen. Im Zentrum steht die Frage, ob und wie die Institute ihre Ziele erreichen. Dabei sind die Ziele und Aufgabenbestimmungen selbst im Kontext innerwissenschaftlicher Entwicklungen und/oder (neuer) gesellschaftlicher Herausforderungen an die Forschung sowie vor dem Hintergrund eines besonderen Legitimationsbedarfs und besonderer wissenschaftlicher Aufgabenstellungen außeruniversitärer Forschung eine Überprüfung zu unterziehen.

Darüber hinaus ist zu spezifizieren, inwieweit in einem Evaluationsverfahren gemeinsame, vergleichbare Bewertungskriterien für die WGL-Institute identifizierbar sind: etwa mit Blick auf einen Forschungstypus, der auf die Verbindung von Grundlagenforschung und gesellschaftlicher Anwendungsorientierung setzt, wenn auch mit unterschiedlicher Schwerpunktsetzung in den Instituten, sowie im Hinblick auf institutionelle Voraussetzungen für spezifische Kooperationsformen in der Forschung sowie für Forschungs"produkte", die auf

langjähriger Forschungserfahrung und Konzeptentwicklung einerseits und Flexibilitäts-/Innovativitätspotentialen andererseits basieren.

3. Die WGL-Institute sind durch ein *komplexes* Aufgaben- und Forschungsfeld gekennzeichnet, das nicht nur durch die kognitive Struktur des Forschungsfeldes bestimmt wird, sondern auch durch unterschiedliche – zum Teil miteinander schwer zu vereinbarende – externe Anforderungen und Kooperationsformen. Für diese Komplexität sowie für die Bearbeitung interdisziplinärer oder multidisziplinärer Forschungsfelder müssen institutionelle und forschungsstrukturelle Lösungen gefunden werden. Hier scheint ebenfalls Evaluationsbedarf zu bestehen. Zu überegen ist jedoch in diesem Zusammenhang die Notwendigkeit einer Ausdifferenzierung von Evaluationsaufgaben in institutioneller/personeller (Evaluatoren) Hinsicht, denn Evaluatoren für die Überprüfung der wissenschaftlichen Qualität der Forschung sind nicht unbedingt auch die Experten für Fragen der organisatorischen Weiterentwicklung des Instituts.

4. Hinsichtlich des zu konzipierenden Evaluationsverfahrens für Blaue Liste-Institute sollte konzeptionell von einem integrierten und komplementären Modell ausgegangen werden, das interne und externe Selbstreflexions- und Begutachtungsprozesse beinhaltet. *Interne Evaluationen* im Sinne von regelmäßigen Selbstverständigungs- und Selbstreflexionsprozessen über den "status quo" und das "quo vadis" sind nicht nur unter dem vorzubereitenden Aspekt externer Forschungsbewertungen zu betrachten, sondern haben ihre eigenständige Bedeutung als systematische Bestandsaufnahme, Leistungs- sowie Stärken- und Schwachstellenanalyse. Nicht auf "Einmaligkeit", sondern auf Regelmäßigkeit angelegt und auf die Erzeugung von Kommunikations- und Einigungsprozessen ausgerichtet, sind sie als Teil von Organisationsentwicklung zu begreifen. In diesem Sinne sollten die bereits eingeführten und etablierten Verfahren in Blaue Liste-Instituten zur Qualitätssicherung (beispielsweise Publication Commitees) sowie die internen Steuerungsgremien hinsichtlich ihrer Funktionalität von den Institutsakteuren einer kontinuierlichen Überprüfung unterzogen werden.

Externe Evaluationen beziehen sich grundsätzlich auf dieselben Gegenstandsbereiche und intendieren eine Überprüfung von Programmen, Zielen und Aufgabenstellungen sowie die Bereitstellung von Expertise für qualitätsfördernde Maßnahmen. Sie unterscheiden sich jedoch insbesondere hinsichtlich des Grades der Differenziertheit beispielsweise der zu erhebenden Daten von internen Evaluationen, da sie sich an andere Adressaten richten. Die Existenz und Ausgestaltung von institutsinternen Maßnahmen zur Qualitätskontrolle und -förderung stellen für externe Evaluierungen ein entscheidendes Bewertungskriterium dar. Die Begutachtungen sind komplementär zur Innensicht der jeweiligen Institute als externe "Bewertungsinstanz" zu verstehen.

5. Mit Blick auf die involvierten Akteure in den verschiedenen Evaluationskreisläufen wird den wissenschaftlichen Beiräten – als "Schnittstelle"

zwischen internen Selbstverständigungsprozessen und externen Bewertungen – seitens der Institute eine besonders große Relevanz als kompetentes und mit dem Institut vertrautes Beratungsgremium beigemessen: Bei der vorgesehenen Übernahme von evaluierenden Aktivitäten durch Institutsbeiräte ist nicht von einer Unvereinbarkeit von Beratung und Bewertung auszugehen, wenn konzeptionell auf eine sorgsame Austarierung unterschiedlicher Aufgaben geachtet wird und Verfahrensregelungen entwickelt werden, die Beiräte nicht in unproduktive Rollenkonflikte mit dem Institut und externen wissenschaftspolitischen Akteuren bringen.

Da der gesellschaftliche Problembezug oder direkte Anwendungsbezug der Forschung bei einem relevanten Teil der Institute das Aufgabenprofil entscheidend prägt, ist die Einbeziehung von kompetenten Praxisvertretern oder "Nutzern" der Forschung ein Desiderat.

6. Um die Akzeptanz der Bewertungen zu erhöhen und die Umsetzung von Empfehlungen zu befördern, sollte die Ergebnisvermittlung von internen und externen Evaluationen als interaktiver Prozeß konzipiert werden: das heißt Bestandteil des Verfahrens sollte der Diskurs zwischen Evaluatoren und Akteuren der Institute über die Ergebnisse der Evaluation sein sowie Expertise bei der Umsetzung von Empfehlungen zur Weiterentwicklung des Instituts zur Verfügung stehen.

Literaturverzeichnis

Altrichter, H., Schratz, M., Pechar, H. (Hg.) (1997), Hochschulen auf dem Prüfstand. Was bringt Evaluation für die Entwicklung von Universitäten und Fachhochschulen? Innsbruck/Wien

Barz, A., Carstensen, D., Reissert, R. (1997), Lehr- und Evaluationsberichte als Instrumente zur Qualitätsförderung. Bestandsaufnahme der aktuellen Praxis, Gütersloh

Hochschulrektorenkonferenz (Hg.) (1998), Evaluation und Qualitätssicherung an den Hochschulen in Deutschland – Stand und Perspektiven, Beiträge zur Hochschulpolitik 6/1998, Bonn

Hochschulrektorenkonferenz (Hg.) (1999), Qualität an Hochschulen. Beiträge zur Hochschulpolitik 1/1999, Bonn

Vroeijenstein, A. I. (1995), Improvement und Accountability: Navigating between Scylla and Charybdis. Guide for External Quality Assessment in Higher Education, London/Bristol

Webler, W.-D. (1998), Das Bielefelder Modell zur Evaluation der Lehre als Organisationsberatung durch Hochschulforscher, in: Evaluation und Qualitätssicherung an den Hochschulen in Deutschland – Stand und Perspektiven, Beiträge zur Hochschulpolitik 6/1998, Bonn, S. 189 ff.

Westerheijden, D.F., Brennan, J., Maassen, P. A. M. (1994), Changing Contexts of Quality Assessment. Recent Trends in West European Higher Education, Utrecht

Jürgen Lüthje

Impulse und mögliche Parameter für die Forschungsevaluation

Die Qualität der Lehre und Forschung an den Universitäten wurde lange Zeit einfach unterstellt. Doch in gleichem Maße, wie das Qualitätsbewußtsein in der Gesellschaft zugenommen hat und der Hochschulbereich expandierte, wurde diese Selbstverständlichkeit in Frage gestellt. Die Antwort wurde in unterschiedlichen Formen der Evaluation gesucht und gefunden: Evaluation an Hochschulen will Qualität prüfen und beurteilen, vor allem die Qualität von Lehre und Studium. In jüngster Zeit hat sich darüber hinaus auch das Bedürfnis entwickelt, ganze Einrichtungen zu evaluieren, auch Forschungseinrichtungen außerhalb des Hochschulbereichs. Im folgenden sollen konkrete Erfahrungen mit der Evaluation von Lehre und Studium dargelegt und in ihrer Relevanz für die Forschungsevaluation betrachtet werden.

Die Universität Hamburg beteiligt sich seit 1994 am Verfahren zur Evaluation von Studienfächern im Verbund norddeutscher Universitäten – ein bisher sehr erfolgreiches Projekt. In diesem Zusammenhang betreibt die Universität allerdings bislang keine spezielle Forschungsevaluation. Vielmehr unternimmt sie den Versuch, die Entwicklung der Universität insgesamt systemisch anzugehen, d. h. nicht auf einige wenige Problembereiche zu beschränken, sondern bei den wichtigsten Faktoren und Elementen der Universitätsentwicklung parallel anzu-setzen.

Zum Hintergrund dieser Aktivitäten gehören staatlich verordnete und sehr einschneidende Sparmaßnahmen (seit 1995 – voraussichtlich bis 2005 – muß jede zweite frei werdende Stelle gestrichen werden; die Zahl des Personals der Universität wird sich dadurch um etwa 20 % verringern). Deshalb hat die Universität Hamburg beschlossen, durch eine externe Beratungsgruppe von zwölf Wissenschaftlerinnen und Wissenschaftlern aus unterschiedlichen Fachgebieten alle Fachbereiche und Einrichtungen begutachten zu lassen. Daraus ergaben sich Empfehlungen für die künftige Entwicklung dieser Fachbereiche und Ein-richtungen. Aufgrund dieser beiden maßgeblichen Faktoren – den Sparauflagen und der Fremdanalyse – können die Ist-Situation und die Entwicklungsperspektiven realistisch umrissen werden.

Hinzu kommt das von der Volkswagenstiftung geförderte "Projekt Universitätsentwicklung" (Pro Uni), mit dem die Universität ihre Selbstverwaltungs-, Verwaltungs- und Organisationsstruktur systemisch optimiert: Im Kontext dieses Projektes wurden ein Leitbild und ein Zielkatalog für die Entwicklung der Universität Hamburg erarbeitet, die in den nächsten Jahren die Grundlage einer strategischen Entwicklungspolitik für die Universität bilden werden. Das Thema "Qualitätsförderung durch Evaluation?" steht im Zusammenhang mit der Evaluation von einzelnen Studienfächern und ist in allen evaluierten Fächern immer wieder auch unter der Fragestellung bearbeitet worden, ob es möglich ist, die Qualität des

Studienangebotes zu bewerten, ohne die Forschung in den entsprechenden Fächern mit zu evaluieren. Wir haben im Sinne eines pragmatischen Kompromisses jeweils den Fächern überlassen, in welchem Maße die Forschung berücksichtigt wird. Praktisch hat das dazu geführt, daß die Forschung in Bezug auf ihre Bedeutung für die Lehre einbezogen wurde.

Eine eigenständige Forschungsevaluation wurde an der Universität Hamburg bisher nicht durchgeführt. Die Evaluation der Studienfächer fand mit Blick auf den Zusammenhang von Lehre und Forschung statt, sie deckt also mitnichten das gesamte Feld der Forschungsevaluation ab. Ich schildere subjektive Eindrücke, die ich aus diesen Verfahren gewonnen habe. Auch die bereits erwähnte externe Beratungsgruppe, die sogenannte Grotemeyer-Kommission, hat mit begrenzten Mitteln versucht, Einschätzungen zur Leistungsfähigkeit und ein Stärken-Schwächen-Profil der Forschung in den einzelnen Fachbereichen zu erarbeiten. Dabei wurde ein nach Fächergruppen unterschiedlicher Mix von Kriterien herangezogen. Im Vordergrund stand die subjektive Beurteilung von Gutachtern als "Peers", basierend auf Informationen, die aufgrund von Gesprächen mit möglichst vielen Fachbereichsmitgliedern verfügbar waren.

Beiden Verfahren, sowohl der Studienfachevaluation als auch der externen Begutachtung, ist gemeinsam, daß sie einen ersten Schritt der Selbstevaluation mit einem zweiten Schritt der externen Bewertung verbinden. Diese Kombination von Selbstevaluation und externer Begutachtung hat sich nach unseren Erfahrungen als ausgesprochen zweckmäßig erwiesen. Die Selbstevaluation wurde mit dem Ziel durchgeführt, zukünftige Entwicklungsmöglichkeiten und Chancen auszuloten. Die externe Begutachtung mußte sich mit der Tragfähigkeit dieser Vision der zukünftigen Entwicklung auseinandersetzen und damit ihrerseits eine Vorstellung zur Zukunft entwickeln.

Als unverzichtbares Element dieser Verfahren haben wir in Modifikation des niederländischen Evaluationsverfahrens ergänzend eine auswertende Konferenz eingeführt: Sie findet am Schluß der Selbstevaluation und der externen Begutachtung statt. In dieser auswertenden Konferenz begegnen sich Evaluierte und Gutachter. Sie können in diesem Rahmen die Feststellungen und Bewertungen noch einmal intensiv in anderthalb Tagen diskutieren. Es hat sich gezeigt, daß vielfach Mißverständnisse und Fehler ausgeräumt werden konnten. Die Bereitschaft der evaluierten Fächer, die Empfehlungen zu übernehmen und praktisch umzusetzen, wurde durch diesen produktiven Dialog erheblich gefördert. Befürchtungen, daß die auswertende Konferenz die Gutachten oder ihren Aussagewert entschärfen könnte und daß die Empfehlungen ihr Profil oder ihren Nachdruck verlören, haben sich nicht bestätigt. Die Gutachten sind auch nach den auswertenden Konferenzen erfreulich klar in ihren Aussagen und gehen sehr offen mit Problembeschreibungen und Schwächenanalysen um.

Dieses Procedere hat aber eine Voraussetzung – und hier komme ich zu einem wichtigen Unterschied unseres Verfahrens im Vergleich mit dem Verfahren für die Institute der Blauen Liste, wie es bisher im Vordergrund der Diskussion gestanden hat: Diese Evaluation stand unter dem Vorzeichen, eine politische Entscheidung über die Existenz sowie die künftige Organisationsform und Finanzierung begrün-den zu müssen. Das ist durchaus legitim, denn wären diese Entscheidungen ohne vorherige Evaluation – gleich welcher Art – getroffen worden, hätte das mit Sicherheit einen Ansatzpunkt für Kritik an der politischen Entscheidung geboten. Es ist unvermeidlich, daß in einem solchen Verfahren schwierige Aporien auf-treten. Wenn eine politische Entscheidung durch ein Evaluationsverfahren auf eine bessere sachliche Grundlage gestellt wird, dominiert die Außensteuerung und die externe Bewertung. In der Universität Hamburg haben wir demgegenüber die Chance gehabt, unsere Verfahren autonom durchzuführen. Die Evaluation verfolgte ausschließlich das Ziel, die Institution mittels Evaluation lernen zu lassen, ihr eine Selbstkorrektur zu ermöglichen.

Ich will nun zunächst ansprechen, welche Zwecke Evaluationsverfahren haben können. Wir unterscheiden zum einen Evaluationen, die dem Ziel des Leistungs-vergleichs oder der Leistungsmessung zwischen Institutionen dienen. Solche Evaluationsverfahren können zwar auch Lerneffekte bewirken, sind aber für eine offene Benennung von Schwächen oder Mängeln nicht sehr günstig. Ein Leistungsvergleich erschwert der evaluierten Einrichtung, mit den ihr selbst bekannten Schwächen und Mängeln offen und ehrlich umzugehen. Ein anderes Ziel der Evaluation kann in der methodischen Begründung von Planungsentscheidungen liegen, z. B. wenn Ministerien über die Existenz eines Studiengangs entscheiden wollen. Ein drittes Ziel verfolgt die Evaluation, wenn sie der Qualitätsentwicklung dienen soll. Wenn dieses Ziel angestrebt wird, muß der Institution, die evaluiert wird, die Angst genommen werden, Schwächen einzugestehen. Denn erst wenn Angstfreiheit und Fehlerfreundlichkeit das Verfahren prägen, kann die Selbstanalyse, bei der die Institution ihr Wissen über eigene Stärken und Schwächen nüchtern und ehrlich darlegen muß, erfolgreich sein und die Ergebnisse einer externen Gutachtergruppe offen und lernbereit zur Diskussion stellen.

Nach unseren Erfahrungen ist es wichtig, bei der Einleitung eines Evaluationsverfahrens genau zu definieren, welchem Zweck es dienen soll. Wenn man Evaluation mit dem Ziel der Qualitätsentwicklung betreiben will, muß man die evaluierte Institution von unmittelbar negativen Folgen freistellen und ihr die Chance zum Lernen geben. Das dürfte in der Forschung nicht anders sein als in der Lehre. In dem Maße, in dem andere Zwecke als die Qualitätsentwicklung mit der Evaluation verfolgt werden, z. B. Entscheidungen über die Mittelverteilung oder aber über die Existenz einer Einrichtung, kann zwar eine Evaluation auch sinnvoll sein, aber es kann nicht erwartet werden, daß aus dieser Evaluation primär und unmittelbar Qualitätsentwicklungs-Impulse hervorgehen. Daß diese Impulse indirekt durch den

Prozeß als solchen doch ermöglicht werden, kann man allerdings nach den Erfahrungen annehmen, die in den vorausgehenden Beiträgen präsentiert wurden.

In der Forschung und in der Lehre kann sich die Evaluation nicht an wenigen Kriterien oder Faktoren ausrichten. Wenn nicht eine Einzelperson evaluiert wird, sondern eine Institution oder Einrichtung, muß die Evaluation unvermeidlich komplexe Zusammenhänge erfassen. Wenn man den gesamten Evaluationsprozeß in Kategorien abzubilden versucht, wie sie in betriebswirtschaftlichen Wertschöpfungsketten üblich sind, wird erkennbar, daß zwischen dem investierten Aufwand (Personal, Sachmittel, Investitionen und – das wird in Hochschulen häufig vergessen – Zeit) sowie dem Ergebnis, das sich nach Quantität und Qualität, aber auch dem geeigneten Zeitpunkt und der Akzeptanz bei den Adressaten messen oder beurteilen läßt, ein Prozeß liegt, der die Leistung hervorbringt.

In diesem Prozeß müssen viele Faktoren zusammenwirken: eine Strategie, welche die Ziele des Prozesses definiert, die funktionale Effizienz, die Zweckmäßigkeit der Organisation, die Intensität der Kommunikation und die Interaktion zwischen den beteiligten Personen. Wenn eine Einrichtung durch Evaluation verbessert werden soll, reicht es nicht, allein den Aufwand und das Ergebnis zu analysieren. Um systemisch zu optimieren, muß die Evaluation die unterschiedlichen Systemfaktoren erfassen. Das gilt auch für die Forschungsevaluation. Wenn sie sich auf eine Einrichtung bezieht, muß sie versuchen, möglichst viele Systemfaktoren zu erfassen und sie in ein optimales Verhältnis zu bringen. Das führt dazu, daß Evaluationsverfahren als Prozesse organisiert werden müssen, nicht als starr geregelte Abläufe – handelt es sich doch um Kommunikationsprozesse, die durch ein hohes Maß an Partizipation geprägt sind sowie durch einen starken Willen, sich über Ziele zu verständigen. Die Verständigung muß vor allem auch die Erwartungen der Adressaten als das eigentliche Qualitätsmerkmal einbeziehen.

In der Forschung kann Qualität am besten an den Erwartungen der *scientific community* als Abnehmer der Forschungsleistungen gemessen werden. Zu einem Evaluationsverfahren gehört aber auch, daß es nicht als Selbstzweck, sondern mit Blick auf Konsequenzen, also die Umsetzung von erarbeiteten Ergebnissen betrieben wird. In unserem Verfahren hat sich bewährt, die Ergebnisse und Empfehlungen durch Zielvereinbarungen zwischen der Leitung der Institution und dem evaluierten Bereich verbindlich festzulegen und in der Folgezeit abzuarbeiten.

Bei der Forschungsevaluation an Universitäten sehe ich das größte Problem darin, daß Universitäten überaus differenzierte Einrichtungen sind. Wenn die Universität Hamburg alle Fächer und Institute evaluieren wollte, müßte sie den gleichen Aufwand betreiben wie die gesamte Gemeinschaft der Blauen Liste-Institute. Mindestens 90 Fachbereiche und Institute bündeln jeweils mehrere Leistungsschwerpunkte. Das größte Problem der Forschungsevaluation scheint mir daher in der Minimierung des Aufwandes zu liegen. Jeder Perfektionismus führt hier in die

Irre. Was an Informationen und objektiven Daten verfügbar ist, sollte vollständig ausgewertet werden. Es wäre jedoch problematisch, in größerem Umfang neue Daten generieren zu wollen. Wenn sich die Daten aus einem eingespielten System, wie etwa dem Publikationssystem der Physik, ergeben, ist das sehr nützlich. Wenn aber in einem höchst komplizierten Publikationssystem wie dem der Rechtswissenschaften versucht würde, ähnliche Indikatoren zu gewinnen, wäre ein Aufwand vonnöten, der wahrscheinlich in keinem Verhältnis zum Nutzen stünde.

Insofern ist auch in der Forschung das sinnvollste und auch praktisch durchführbare Verfahren die Kombination einer Selbstanalyse mit einem externen Begutachtungsverfahren, in dem beide Seiten versuchen, alle verfügbaren Informationen und Daten offen zu legen und mit diesem Datenbestand pragmatisch und subjektiv beur-teilend umzugehen. Objektivität wird man dabei nur in engen Grenzen erreichen.

Völlig unvermeidlich ist, daß bei der Forschungsevaluation zunächst nach disziplinären Kategorien gewertet wird. Interdisziplinarität kann zusätzlich in die Bewertung einbezogen werden, aber die Basis wird eine disziplinäre Bewertung sein. Sie kann zunächst einmal im nationalen Rahmen vergleichen, muß aber versuchen, internationale Maßstäbe einzubeziehen.

Darüber hinaus ist bei der Forschungsevaluation sehr wichtig, auch nach den Forschungsvoraussetzungen zu fragen: Ist die Organisations- und Infrastruktur tatsächlich leistungsfördernd? Sind die festgestellten Ergebnisse diesen Voraussetzungen angemessen? Auf der Grundlage einer schwachen und unzureichenden Struktur können beachtliche Forschungsergebnisse erreicht werden, wenn gut geforscht wird. Umgekehrt kann ein Ergebnis mit Riesenaufwand zustande gekommen sein, der sich dann nur durch Exzellenz oder besonders herausragende Qualität rechtfertigen läßt. Man sollte also die Forschungsleistung nicht für sich betrachten, sondern ihre Einbettung in eine gegebene Organisations- und Infra-struktur berücksichtigen.

Die Evaluation der Forschung weist gegenüber der Evaluation von Lehre und Studium wichtige Besonderheiten auf, die bei der Gestaltung und Handhabung des Verfahrens zu berücksichtigen sind. Der wichtigste Unterschied dürfte in dem wesentlich größeren und offeneren Adressatenkreis und in der vergleichsweise noch höheren Komplexität und Spezialisierung der Forschung liegen. Insbesondere im Hinblick auf die Qualitätsbeurteilung stellt darum die Forschungsevaluation noch höhere Anforderungen. Dennoch erscheint es möglich und zweckmäßig, bei der Entwicklung von Verfahren der Forschungsevaluation auf die Erfahrungen mit der Evaluation von Lehre und Studium zurückzugreifen. Insbesondere die Kombination von Selbstanalyse und externer Begutachtung dürfte sich auch in der Forschung bewähren. Ebenso wichtig ist die Klärung der Ziele und Zwecke der Evaluation. Qualitätsentwicklung kann durch autonome Verfahren besser gefördert werden, während der Leistungsvergleich oder die Leistungsmessung eher einer externen

Verfahrenssteuerung bedürfen. In beiden Varianten der Evaluation ist der Kommunikation zwischen Evaluierten und Evaluierenden besondere Bedeutung zuzumessen.

Literaturverzeichnis

Lüthje, Jürgen (1997), Verfahren und Elemente systemischer Qualitätsentwicklung. Beispiele aus der Universität Hamburg, in: Stifterverband für die Deutsche Wissenschaft (Hg.), Qualitätsentwicklung in einem differenzierten Hochschulsystem. Dokumentation eines Symposiums, Wissenschaftszentrum Bonn, 9. Januar 1997, Essen

Fischer-Bluhm, Karin (1998), Evaluation im Verbund Norddeutscher Universitäten, in: Hochschulrektorenkonferenz (Hg.), Evaluation und Qualitätssicherung an den Hochschulen in Deutschland – Stand und Perspektiven. Nationales Expertenseminar der Hochschulrektorenkonferenz, Bonn, 29. Mai 1998, Beiträge zur Hochschulpolitik 6/1998

Adrian C. L. Verkleij[20]

Self-evaluation and External Review

Introduction

This contribution to the Seminar on Quality Improvement by Evaluation is based on my personnel reflection on my experiences as programme manager of the Dutch university research assessment system, and on my work as consultant for the Center of Higher Education Policy Studies (CHEPS) in the field of quality management in higher education in the Netherlands as well as abroad. The aim of this contribution is to share the Dutch experiences with self-evaluation in universities. Self-evaluation is generally seen as an effective instrument for institutional change (Mets 1997) , but is it is also stated that any self-evaluation should be followed by a form of external review to add rigor to the outcomes of the self-evaluation. I will especially focus on the interaction between internal and external assessments, because the way external assessments are organised, has a considerable influence on the way self-evaluations are carried out. All actors in evaluation processes act strategically, taking into account the purpose and the intended follow-up activities (decision context) of the evaluation exercise. I conclude that strategic behaviour is part and parcel of any evaluation process, and that one should learn to live with it. Outcomes of an evaluation process ought to be seen as input for a decision making process. They can not replace decision making processes as such. They can deliver semi-objective, semi-comparable data and as such they can contribute to the transparency and the acceptance of decisions, but in the end someone has to take these decisions, taking into account all other factors that are relevant in a specific context.

The Dutch University Research Assessment System

I will not present a comprehensive description of all developments with respect to research assessment and research policy in the Netherlands. Others have done that recently (Rip and Van der Meulen 1995,Van der Meulen and Rip 1998). I will concentrate on what we have learned from our experience with external research assessments of university research on a national scale.

Since 1982, every five or six years all university research has been assessed by external and independent international peer committees. The purposes of this nation-wide system changed over time, as did the central actors in the system (Fig. 1). This illustrates that external assessment systems are not stable over time but show

[20]Adrian Verkleij is Senior Advisor at CHEPS. In the period 1994-97 he worked for the VNSU where he was responsible for the university research assessment system. His address is: P.O. Box 217, 5300 AE Enschede, The Netherlands, E-mail: a.c.l.verkleij@cheps.utwente.nl

developments, due to: (a) internal learning experiences and (b) changing external circumstances

Figure 1: Developments in the Dutch university research assessment system

Period	**Actor**	**Goals**
1982 - 1987	Government	Budget allocation based on ex ante appraisals of newly formed research programs
1987 - 1992	Government and VSNU	Accountability, based on ex post assessments of research programs
1993 - 1997	VSNU	(1) Improvement and (2) accountability based on self-evaluation (ex post and ex ante)
1997 - 2002	VSNU, individual universities and institutes	(1) Improvement, (2) mission oriented (3) input for other processes, including accountability
2002 - 2007 ?	Individual universities and institutes ?	(1) Part of a "quality culture" (2) integrated in strategic planning

Changing responsibilities for assessing university research

The initiative for the introduction of a national research assessment exercise goes back to the beginning of the nineteen-eighties, when the Minister of Education and Science imposed a national peer assessment system for research "*voorwaardelijke financiering*" (conditional financing) on the universities. At that time the major challenges were: (a) to change the university research organisation from an individualistic to a programmatic base, and (b) to find a way to connect quality of research with funding. Universities were forced to define research programmes with a minimal research input of five full-time equivalents of staff research time. Ex-ante appraisals were organised by the Ministry, in co-operation with the Royal Academy of Arts and Sciences, the Dutch Organisation for Fundamental Scientific Research (ZWO[21]), the Royal Institute of Engineers (KIVI), and some other more specialised advisory councils. After 3 to 4 years it became clear that all university research could be "programmed" and that a vast majority of the university research programmes

[21] Later its name changed into The Dutch Organisation for Scientific Research (NWO)

received a "plus" on a two points scale (the other score was a minus, which meant that research quality was below standard). It also became clear that this exercise with research assessments did not lead to a solid base for interuniversity reallocations of research budgets.

In the second round of *voorwaardelijke financiering,* which started in 1987 some major changes were negotiated between the newly founded Dutch Association of Universities (VSNU) and the Ministry. The initiative to (re)design the conditional financing system came from the VSNU and the negotiations with the government were based on a VSNU proposal for this second round. Although inter-university reallocation of budgets was still at stake, these negotiations never lead to an agreement on a (re)allocation mechanism. As programs had been formed in the first round, it became possible to assess outcomes of research programs (ex-post assessments) in the second round and thus, the emphasis shifted from ex ante to ex post assessments. This shift was seen by the universities as a further acknowledgement of their autonomy. Ex-post assessment were seen as a way to organise accountability, and as a way to minimise state influence in university research programming. This emphasis on 'autonomy' and accountability, also lead to a situation where in which no further discrimination in quality standards other than "plus" or "minus" were accepted. The VSNU became responsible for the process management of the second round of conditional financing. The Royal Academy, the Dutch Organisation for Scientific Research and the Royal Institute of Engineers were made responsible by the Ministry for the actual assessments of the research programmes.

In the third round, which started in 1993, the idea of budget reallocation between universities was dropped completely[22]. The name '*voorwaardelijke financiering*" was dropped. The new law on higher education made the universities responsible for quality assurance mechanism for both teaching and research and the universities assigned this task to the VSNU. The VSNU (VSNU 1994) redefined the system. The major aim of the quality assurance procedure became quality improvement. Accountability was seen as a by-product, which was served by making the assessment reports public. It was also intended that the VSNU research assessments should be used as input for other processes like research foresight committees and the accreditation of (inter-) university research schools by the Royal Academy. During this period some universities started to discuss that individual universities should organise their own research assessments, basically because they saw it as an essential element in internal institutional quality assurance mechanisms, and also because they wanted to have more detailed and more contextually organised feedback than the VSNU could deliver, and it was felt that this detailed information should not necessarily be made public. Research assessments were seen as an

[22] Internal follow up procedures which may include internal budget decisions, and all other kinds of effects are observed frequently (Westerheijden 1997).

essential ingredient of internal strategic processes in a competitive environment. A number of universities (or faculties or institutes within universities), started to organise their own peer assessments, a policy that was stimulated by the Royal Academy as part of the re-accreditation processes of research schools. In some cases they invited the VSNU to organise these external assessments.

In 1998 the universities decided that the VSNU should organise the fourth round of research assessments to be based on the – mainly good- experiences with the former one organised by the VSNU, but once again some essentials changes were made. This round will again encompass all university research disciplines and will take another 5 years. It is hard to predict if their will be a fifth round, starting in 2003. If I may speculate, I guess that the need of external feedback by competent peers or experts as part of strategic planning in a more market-oriented context will increase and that this will lead to a more individualistic and more contextual approach. If I am right, then this may greatly challenge the collective approach followed by the VSNU up until now. However, this workshop is not the right place to discuss the future of the Dutch research assessment system.

Conclusion

In the Netherlands the goals changed over the years from funding (like it is still the case in the UK research assessment exercises), but which failed in the Netherlands, to accountability (which is an important goal, but the added value of ex-post evaluations for improvement processes as such is limited), and then to quality improvement (by organising feed back from well respected peers and experts on both pas -performance and future outlooks), which is highly appreciated and effective (Westerheijden 1997).

Over the course of the different assessments, responsibility for external quality assurance for university research shifted from the Ministry of Education and Science in the first round, to a shared responsibility of the Ministry and the collective body of the universities (the VSNU) in the second, and then further on to the VSNU in the third and forth round. Parallel to the national research assessment system, universities, faculties and institutes started to organise their own – more dedicated – external assessments as input for their own policy and management processes. In addition the outcomes of these external assessments were used as input for other procedures, like fund raising, accreditation of research schools, etc. More than 15 years of experience with external quality assessments has lead to the emergence of a quality culture in which regular evaluations have become an intrinsic part of academic life. As a consequence, a national assessment system directed to “product” assessments may no longer be needed and therefore the character of a fifth round may have to be radically changed.

Learning Experience

After this short description of developments of the Dutch system for university research assessments, I would now like to turn to some reflections, based on my experience as co-designer and implementers of the VSNU research assessments in the period 1993-1997.

Most of the quality assurance systems I am familiar with are based on an interaction between self-evaluation followed by an external review. Both processes are organised in a context which influences both the behaviour of the unit (institute, faculty, university) under assessment as well as of the external assessor. An important aspect of this context is how the assessments are followed-up, or in other words, who does what with the outcomes of the external assessment. In my experience the importance of clear statements concerning this is often underestimated. Follow-up procedures are often vaguely described and insufficiently communicated, which often leads to a situation in which different parties in the assessment process have different perceptions of the follow-up. The consequence is that each party starts to speculate about consequences and follow-up procedures, and this may lead to uncertainty and to a defensive attitude. Simple questions as: what is going to happen if a unit scores excellent, and what if unsatisfactory? Does it matter at all? If yes, how? These questions should be raised, answered and communicated in advance, as part of the design process of an assessment system. Clear answers are necessary in order to create open minded and co-operative attitudes which are needed for a honest self-evaluation. Trying to answer these questions may also be useful as a step in the design process. It prevents you from developing assessment instruments without knowing which problems you wish to solve.

My experience is that the consequences of the assessment outcomes influence the behaviour of both the unit under scrutiny and the external assessor. If external consequences (especially funding) are dominant, each unit will go for maximum scores and will hide weaknesses. Open and constructive self-evaluation reports will not be produced. Units may present short-time solutions, compliance and lip-service. External assessment committees are seen as belonging to “the enemy” and are treated as such.

Members of external assessment committees have their own loyalties and they have to choose who they are loyal to. They have several options: their commissioner, their scientific discipline, their professional group or the unit under assessment. “Why should we be critical if the result may be that funds reserved for History may shift to Physics”, one member of the VSNU Committee for Archaeology and Historical Sciences in the Netherlands complained. Their (often implicit) choice may influence their behaviour, and this may lead to strategic reports, rather than giving an open overview of strengths and weaknesses.

If internal goals, like quality improvement are the major aim, one may observe a shared interest between the unit to be assessed and its external assessors. In such cases self-evaluation can be used to identify strengths and weaknesses realistically, and units may be challenged to present plans and strategies to overcome the weaknesses that came out of the self-evaluation.

The external committee then may concentrate on assessing the outcomes of the self-evaluation process as well as the future plans. Being critically constructive and constructively critical then becomes the major intention. I often observed signs of relief from the Members of the Peer Committees, as I explained to them that they were expected to be supportive by giving advice to the units' management and that the management of the unit, or its supervisors (deans, rectors), and not the government are responsible for follow-up activities.

Conclusion

Thus, in each context, both parties (the institute under scrutiny and the external assessor) act strategically. Often professionals learn fast how to manipulate an assessment system, especially when they expect it to be hostile to them. External assessors make (often implicit) choices who they are loyal to. Evaluation processes therefore must be characterised as political processes. The frequently made claim that these processes lead to "objective" results therefore is of relative value.

Personally, I do not have problems with a weakened claim for "objectivity" as long as the users of the outcomes of such processes are aware of the context in which the assessment exercise was organised. Users must be able to read between the lines of the Report and they should be able to understand the report in its context.

Definition of quality

An important aspect for any (self-)evaluation is the definition of quality. Nowadays may people accept that "quality" is a broad term which has different meanings in different contexts.

However, among university professors the traditional and dominant view is that quality of research relates directly to the "contribution to the international progress of science". There is a lot of truth in this, but nevertheless this statement has become highly debated in the Dutch assessment system (Verkleij 1999, Verkleij and Huisman 1999).

It is a debate between the classical universities on one hand and the universities of technology on the other hand. In the course of the last round of assessments the technical universities blamed the VSNU assessment committees for using quality and

productivity criteria derived from fundamental science to assess technical sciences and therefore for neglecting the special characteristics of technology.

The notion of quality was also debated in some humanities and social science disciplines. Some groups claimed that these sciences should escape from Dutch parochialism and should participate as members of the international scientific community. These groups argued that all research units should be assessed based only, or mainly, on international publications. Other units claimed that their mission was to serve and support Dutch culture and professional groups in the Netherlands. They believed that they could only serve these groups by publishing reports, books and articles in Dutch.

Nowadays around one third to even 40 percent of the university research budgets in the Netherlands come from third parties. The nature of this "commissioned research" varies from fundamental work to applied research, but in general, applied oriented or strategic research is dominant. However, the university research groups which bring significant amounts of additional money to universities are often assessed along criteria derived from fundamental research, leading to significant lower scores (and prestige).

The VSNU took these critics seriously. It appeared that the basic question behind the discussion of quality definitions was whether or nor differences in research missions between universities, departments, units and even individual researchers should be taken into account when assessing research quality.

Most Dutch universities and faculties nowadays have mixed research portfolios, covering a whole range of activities, varying from fundamental research to applied research and research based consultancy. Each of these activities has its own criteria for quality.

Thus, in the present round, which started in the second half of 1999, faculties and departments are challenged to define "mission statements" which should indicate: (a) the scope of their research, (b) their ambitions in their research field, (c) their clients or audiences and (4) their publication policy to address their clients or audiences.

To answer these questions, a unit really has to go through a self-evaluation process. Looking at a whole university one may see that research missions vary from faculty to faculty, from department to department, and that they may even vary within departments.

Quality then becomes related to "fitness for purpose". These notions of quality as "fitness for purpose" and mission based assessments have now been introduced in the Dutch research assessment system.

Self-evaluation

Most self-evaluation reports, that I have seen so far in the Netherlands (and in other countries) tend to be more descriptive than evaluative. The VSNU Protocol for the research assessments (VSNU 1998) speaks about self-evaluation, but it does not give any instructions for the way self-evaluation should be carried out.

In our consultancy activities, which are often directed towards introducing quality improvement processes, we use a model for self-evaluation that consists of three steps (Verkleij 1997, Reinecke 1998).

(a) Defining the institute's mission

The VSNU Protocol 1998 suggest that a mission statement covers:

- The scope/area of the research
- The nature of the work (in terms of applied ←→ pure , curiosity ← → society driven and monodisciplinary ← → multi-interdisciplinary)
- The objectives and ambition (envisaged achievements)
- The audience (or clients) (academic community, professional audiences, society, students).

Internal discussions concerning finding the present and the desired balance between these four aspects are both challenging and rewarding parts of a self-evaluating exercise. It is important to accept diversity. Different faculties or departments may define different missions.

(b) A SWOT analysis identifying *internal* strengths and weaknesses and *external* opportunities and threats. In a SWOT analysis faculties and institutes are challenged to define their own quality criteria and quality standards and to apply them to their own situation. Important elements in a SWOT analysis are:

- relation to the present and *future* mission of the institute or organisation
- scientific position in relation to internal and external scientific developments
- present and future position in addressing societal developments (including funding opportunities)
- quality of staff (including leadership) but also i.e. mobility
- quality of facilities
- present and future financial resources
- position in scholarly networks
- position in professional networks
- innovation capacity, etc

We often suggest to use scoring techniques to stimulate institutions to define clear statements about each of these aspects, but the arguments used to formulate such a score often are more important than the scores themselves.

(c) The development of a future strategy or strategies that build on strengths and opportunities, aimed at overcoming weaknesses and to 'managing' threats.

These three elements form the core of any self-evaluation report about research. In our view each self-evaluation should be followed up by some form of external assessment. The major assignment of the external assessor then becomes whether or not the promises made in the mission statement have been realised (in case of ex-post assessments) or probably can and will be realised (in case of an ex-ante appraisal). The input for the external assessment committee comprises the three elements mentioned before: the mission statement, the outcomes of the SWOT analysis and the future strategy.

The external assessment is formative by nature. It not only provides independent assessments, but also gives recommendations. Theoretically, the external assessment could be limited to the assessment of the self-evaluation process, but in practice discussions about the outcomes (or the choices made or the quality delivered) may also occur.

Conclusion

The introduction of self-evaluation processes is an important step forward in the direction of a mature assessment system, and an important step in the development of a quality culture. If prepared and implemented well, self-evaluation processes may lead to a maximum of learning experiences. However, the downside may be that outcomes may be rather soft in terms of accountability. Thus, a necessary prerequisite is that a national government, or in case of the Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz, the Board accepts such an approach.

Use of performance indicators

The Dutch assessment system is not based on performance indicators, but some basic quantitative input and output data are used. My experience with the use of performance indicators for assessing research performance has taught me that the outcomes are diverse.

A few observations

(1) Discussions about the relevance of performance indicators are often very emotional, because behind the selection of indicators lies the question about the acceptance of the diversification of research missions and somebody's individual contributions to those missions. Persons involved in fundamental science often claim that their science, by definition is "better" than applied research. Often "applied scientists" deep in their heart, support this claim, because that was the way they have been socialised by the academic community. However, they notice that at the same time, the risk to be rated as second class scientist.

(2) What do science indicators really measure? Most science indicators measure "impact". The VSNU Committee that assessed Physics (VSNU 1996) stated that citation indices, measured relevance more than quality. To them "quality" is in the ideas, in creativity, in the "unexpected". In general, science indicators measure the contribution to the international, English speaking, scientific community. As stated earlier, this is a relevant aspect for those research programmes that consider "contributing to the international forefront of scientific development" as their major aim. Especially in organisations for applied research, but also in universities where more and more contract-research is carried out, one should be aware that applied researchers have other ways of disseminating their research findings than publishing in highly-rated scientific journals. When I was involved in research assessment exercises I participated in serious discussions about the use of bibliometric methods. In the natural sciences there were different opinions due to the different position of the technology orientated programmes in this field. In the social sciences, in which those groups that, after a long struggle had realised a number of international publications wanted to receive the credits for that, while the more practice oriented research groups, which publish a lot in Dutch were in favour of a more inter-subjective approach of the Review Committees. And, in the humanities some of the assessment committee saw publications in Dutch as equal important as international accessible publications, due to their mission, to support knowledge of the national heritage.

(3) Almost all VSNU Review Committees used some kind of input-output comparisons. However, I often advised the Committees to use these comparisons only for internal purposes. These comparisons provide some first indication of topics that may be questioned, and need further scrutiny. Most peer committees are able to use these parameters as part of their more holistic approach of research performance. But, to my experience, as soon as these data are officially published, they lead to huge discussions about the correctness, the reliability and the relevance of these data, distracting time and energy from the real issue, which is the assessment outcome itself.

(4) Often members of review committee itself are very suspicious about the use of quantitative data, because they see them as part of bureaucratic processes which neglect their abilities as being "peers" who are able to produce their own independent views;

(5) The last point I would like to mention is that some other peer assessment committees, may tend to hide behind quantitative data instead of producing their own, more holistic view about the quality of research, because they lack courage to rely on their own subjective or intersubjective view of the programmes.

Conclusion:

I am not against the use of performance indicators, as long as:

- They are used as input and not as the result of an assessment exercise;
- They are linked to the purpose/mission of the research programmes;
- They are used by persons who are well aware of the limitation of bibliometric data.

Acceptance of assessment systems

The Dutch system for university research assessment has been designed mainly to improve the quality of research. This means that in the end the system intends to change the behaviour of academics on the shop floor. Therefore the VSNU looked carefully into factors that might enhance the acceptance of the assessments given, because it believes that acceptance is the key factor to prolonged and sustainable effects.

Key words for the design of the system therefore became:

- Involvement: of the researchers or their representatives (Deans of Faculty) in the design of the process;
- Diversity: within the framework of a general protocol, the deans of the faculties involved are invited to specify the relevant characteristic of their discipline as input for the assessment procedure
- Openesse: the prescriptions for self-evaluation processes offer researchers ample room to present themselves in the way they like
- Trust: deans of faculties can specify the qualities of review committee members. They may suggest names for the chair and the members of the assessment

committees, who are appointed by the President of the VSNU after consultation of the Royal Academy of Arts and Sciences, in order to guarantee independence.

- Transparency of the process: although a peer assessment, by definition, has some characteristics of a “black box” the transparency of the process is greatly increased by organising face-to-face interviews with the programme directors involved, not only for information and verification, but also because it gives the researchers the opportunity to check the "quality" of their peers (it is a two-way process).
- Transparency of the outcomes: to enhance the transparency of the outcomes, the VSNU asks each review committee: (1) to specify the way they operationalised the four assessment aspects used (scientific quality, scientific productivity, scientific and societal relevance and viability), (2) to use an assessment scale (a five points scale for each of these aspects) and (3) to add a few lines of explanatory texts to each assessment (these may include formative aspects).
- Transparency of the follow up: It should be made clear in advance who is going to do what with the outcomes of the assessment, and one should stick to the defined outcomes. However, one should realise that the way the outcomes are used may differ from university to university and that the outcomes of an external assessment process only are one parameter in an often very complex process of institutional decision-making.
- The rigor of the design: each assessment system can and will be manipulated by the researchers, administrators, committee members, etc. Based on experience gained, one should act proactively where possible during the design of the process and one should be aware of strategic behaviour without making the design overly bureaucratic.
- Communication and Training: assessment systems are described on paper. Our experience is that these papers are not always read carefully and may be interpreted in different ways. Therefore the researchers being assessed and the management responsible for the preparation of the assessments often feel insecure about what is expected from them (and how they can use/manipulate the system effectively). Therefore, workshops for researchers and managers involved are very helpful and enable the participants to discuss the design of the system, including its intended effects. I experienced that especially “stories” of similar institutes that went through such processes before, were extremely helpful.

Summary

In summary, there are four major conclusions that can be drawn from our experience with research assessments:

- Define the decision context in advance and inform participants about the decision context.
- Be aware of the strategic behaviour of every participant in an assessment process and find ways to live with it
- Try to define common goals and interests among the many groups involved – including the scholars to be assessed, local management of the institutes and top management of the system (including Ministries), as a basis for self-evaluation processes.
- Make use of existing experience with research assessments to design and implement self-evaluation processes.

Literature

Mets L. A. (1997), Planning Change through Program Review, in: Peterson Dill, Mets and Associates (Eds.), Planning an Management for a Changing Environment, San Francisco

Reinecke, C. J. (1998), Research Policy as an integral Part of a University's Strategic Planning, SAUVCA Publication Series, 98/3, p. 47-59

Rip, A. / van der Meulen, B. J. R. (1995), The Patchwork of the Dutch Evaluation System, in: Research Evaluation 5 (1), p. 45

Van der Meulen, B. and Rip, A. (1998), Mediation in the Dutch science system, in: Research Policy, 27, p. 757-769

Verkleij A. C. L. (1997), Policy Document for Quality Promotion of Postgraduate Education and Research, Potchefstroom University for Christian Higher Education, South Africa

Verkleij A. C. L. (1999), Different approaches to defining research quality, in: News for the Human Sciences, Vol. 5, No. 2

Verkleij A. C. L., Huisman, J. (1999), Fundamenteel of toegepast, een vruchteloze discussie, Interisciplinair, 10(2) (in Dutch)

VSNU (1994), Protocol for the Research Assessments, Utrecht

VSNU (1996), An Analysis of Physics in the Dutch Universities in the Nineties, Utrecht

VSNU (1998), Protocol for the Research Assessments, Utrecht

Westerheijden, D. (1997), A solid base for decisions, in: Higher Education 33, p. 397

Ulrich Teichler

Hochschulevaluation und Hochschulmanagement im internationalen Vergleich – einige Thesen

1. Zur Evaluation von Forschungsinstituten und Hochschulen

Die Bedingungen für Evaluation von Forschungsinstituten und Hochschulen unterscheiden sich dramatisch. Übersicht 1 stellt die Bedingungen idealtypisch-kontrastierend dar.

Neun Unterschiede lassen sich vor allem nennen:

(1) In den meisten Ländern gab es an Universitäten keine Tradition der Evaluation, verstanden als systematische regelmäßige und flächendeckende Einschätzung ihrer Leistungen (z. B. aller Wissenschaftler des gleichen Institutionstyps). In Deutschland wurden die einmal berufenen Professoren in der Regel nur dann in ihren Leistungen bewertet, wenn sie dazu selbst besondere Anlässe gaben: wenn sie zusätzliche Forschungsmittel einwarben, ein Manuskript zu Publikationszwecken einreichten oder sich auf eine andere Professur bewarben. An den Forschungsinstituten in Deutschland war es jedoch bereits seit langem üblich, die Leistungen regelmäßig durch Beiräte beobachten zu lassen und nach längeren Zeitspannen eine Inspektion unter der Fragestellung vorzunehmen, ob das Institut fortgeführt werden sollte; bei der Max-Planck-Gesellschaft zum Beispiel stand eine solche Prüfung immer kurz vor der Emeritierung des Direktors an.

(2) Die Ziele eines Forschungsinstituts sind in der Regel homogen, die einer Hochschule dagegen möglicherweise plural und heterogen. Während ein Forschungsinstitut ausschließlich oder primär für die Forschung zuständig ist, kommen bei den Hochschulen Lehre und Studium und eventuell Dienstleistungen hinzu, die mit Forschung und Lehre verbunden sind. Dies hat unter Umständen zur Folge, daß Schwächen einer Funktion durch die Bedeutsamkeit im Hinblick auf eine andere Funktion ausgeglichen werden können. Mediokrität der Forschung stellt die weitere Existenz eines Forschungsinstituts in Frage, mag aber bei einer Hochschule, die eine abgelegene Region mit Studienangeboten versorgt, akzeptabel sein. Die Umfragen der SPIEGEL-Studie von 1999 lassen zum Beispiel den Schluß zu, daß an deutschen Universitäten eine negative Korrelation von Forschungsqualität und akzeptablen Studienbedingungen besteht.

(3) Damit hängt zusammen, daß es eine deutliche horizontale Systemdifferenzierung von Hochschulen des gleichen Typs geben kann, dagegen bei den Forschungsinstituten kaum. Ersteres war in Deutschland in der Vergangenheit selten der Fall, mag jedoch in Zukunft zunehmen, wie die weitverbreiteten Forderungen nach einer Profilbildung der Hochschulen bzw. Fachbereiche zeigen. Die einzelnen Forschungs-

institute haben sich dagegen im Bereich ihrer inhaltlichen Akzentsetzung oft mit anderen Forschungsinstituten zu messen, deren Profile sich nur bedingt unterscheiden.

(4) Auch die Forschungsfunktion der Hochschulen ist vielfältiger. Die Forschungsinstitute werden in erster Linie nach der Forschungsleistung beurteilt. An den Hochschulen geht es, wenn wir von Forschung reden, daneben oft um die Qualifizierung des Nachwuchses oder um die Auseinandersetzung mit dem wissenschaftlichen Fortschritt, um die Lehre auf den neuesten Stand zu bringen.

(5) Der größeren Homogenität der Funktionen entspricht es, daß die an der Evaluation von Forschung beteiligten Evaluatoren in ihrer Herkunft und Kompetenz gewöhnlich relativ homogen sind. Erfolgt dagegen Evaluation an den Hochschulen, können sehr unterschiedliche Personen evaluative Funktionen haben: so z. B. mögen Evaluationen an den Hochschulen ganz aus der Sicht der Studierenden oder ganz aus der Sicht von Professionsverbänden erfolgen.

(6) Evaluationen von Forschungsinstituten geht zumeist von der Frage aus, ob die Forschung in dem Institut eine hohe wissenschaftliche Qualität verspricht. Bei der Evaluation von Hochschulen geht es dagegen nicht notwendigerweise um Spitzenleistungen: Es kann primär zur Überprüfung anstehen, ob die Institution ihre selbst- oder fremdgesetzten Ziele konsequent verfolgt, ob sie bestimmten Mindestansprüchen genügt oder welches die am wenigsten wissenschaftlich ertragreichen Bereiche sind, um diese zu mobilisieren oder zu reduzieren.

(7) Selbst wenn wissenschaftliche Qualität nicht als direkt planbar gilt, gehört es zu den Grundannahmen der Evaluation von Forschungsinstituten, daß die Institution durch strategische Entscheidungen – zum Beispiel durch relativ rasche Verschiebung der inhaltlichen Akzente – bedeutsame Beiträge zum Erfolg der Forschung leisten kann. Im Hinblick auf Hochschulen wird eine solche Vorstellung weitaus kontroverser eingeschätzt.

(8) Eng damit verbunden wird für Forschungsinstitutionen durchgängig angenommen, daß eine gewisse Betriebsförmigkeit und eine Kooperation der Wissenschaftler der Forschungsleistung nützt und daher als Rahmen vorgegeben werden kann. Bei den Hochschulen steht jedoch solchen Vorstellungen die weitverbreitete Überzeugung entgegen, daß eine "organized anarchy" den Leistungen der Hochschulen angemessen sein mag.

(9) Das Sanktionspotential ist bei der Evaluation von Forschungsinstituten gewöhnlich höher als bei Hochschulen. Eine Hochschule mag aufgefordert werden, ihre Akzentsetzung zu modifizieren; sie mag auch einen Teil ihren Ressourcen verlieren. Bei Forschungsinstituten dagegen sind Schließungen, deutliche Reduzierungen der Ressourcen, größere Veränderungen bzw. erheblicher Ausbau nicht selten Gegenstand der Empfehlung.

Übersicht 1: Unterschiede in den Bedingungen von Evaluation an Forschungsinstituten

Dimension	Forschungsinstitute	Hochschulen
(1) Tradition	Beiräte, Inspektions-Evaluation	Keine Evaluation (evtl. anlaßbedingt bei positiven Sanktionen)
(2) Ziele der Organisation	Homogen	Plural, heterogen
(3) Horizontale Systemdifferenzierung	Gering	Möglicherweise hoch
(4) Konfiguration der Forschungsfunktionen	Begrenzte Zahl von Funktionen (Qualität/Relevanz)	Multiple Funktionen (z.B. Fitneß der Lehrenden, Nachwuchsförderung)
(5) Komposition der Evaluatoren	Relativ homogen	Heterogen
(6) Zentrale Ebene der Bewertung	Spitzenleistungen	Verschieden (oft Mobilisierung der unteren Ebene)
(7) Strategiefähigkeit der Organisation	Mittel	Gering
(8) Betriebsförmigkeit und Kooperationsgebot	Mittel	Gering
(9) Sanktionspotential	Institutionelle Diskontinuität	Institutionelle Modifikation

2. Zur Einführung von Evaluationssystemen

Vor zwei Jahrzehnten gab es allenfalls in den USA ein Hochschulevaluationssystem. Inzwischen gibt es Hochschulevaluationssysteme in ungefähr zwei Dutzend Ländern. In Westeuropa gehört Deutschland zu den Spätstartern.

Die Hochschulevaluationssysteme wurden in den letzten zwei Jahrzehnten unter ähnlichen Bedingungen des Wandels eingeführt:

(1) Relative Kostensenkungen: Zumeist wurden zugleich entweder die Etats der Hochschulen gekürzt oder es wurde von ihnen erwartet, kostenneutral mehr Studierende aufzunehmen und möglicherweise die Forschungsleistungen unter solchen Bedingungen zu halten oder zu verbessern.

(2) Steigender Managerialismus: Zumeist wurden die staatliche Kontrolle der innerhochschulischen Prozesse verringert, die Position der Hochschulleitung gestärkt und von den Hochschulen eine zielorientierte Institutionsstrategie erwartet.

(3) Erhöhter Relevanzdruck: Die Hochschulen wurden verstärkt der Erwartung ausgesetzt, ihre gesellschaftliche Nützlichkeit in der praktischen Relevanz der Forschung sowie in einer berufsnützlichen Qualifizierung der Studierenden nachzuweisen.

Alle diese Veränderungen sind nicht günstig für die Entwicklung einer Evaluationskultur, die sich von wissenschaftlicher Selbstregulation und externem Vertrauen einerseits und externer Kontrolle durch Externe basierend auf Mißtrauen andererseits dadurch unterscheidet, daß ein weitgehender Normenkonsens von externen Erwartungen und internen Überzeugungen besteht und daß die Evaluierten auf dieser Basis zum Eingeständnis von Schwächen und zur Besserung bereit sind.

Die Einführung von Evaluationen an Hochschulen wird gegenüber der früheren Situation einer okkasionellen Leistungsbewertung zunächst als ein erheblicher Schritt zu externer Intervention verstanden. Da aber Evaluation als interne Intervention nur bedingt funktionieren kann, wenn die Wissenschaftler die Kooperation versagen, wurde bei der Einführung von Hochschulevaluationen der Kompromiß gewählt, daß "Qualität" als höchstes Ziel gepriesen und den Wissenschaftlern selbst die Bestimmung der Evaluationskriterien und die wissenschaftliche Bewertung ihrer "Peers" überlassen wurden. Keineswegs läßt sich prognostizieren, ob dies langfristig so sein wird, denn zu den Zwecken der Einführung von Hochschulevaluationen gehörte auch, die Relevanz der Hochschulaktivitäten und deren Effizienz zu steigern – Kriterienbündel, die bei der Bewertung durch wissenschaftliche Peers nicht notwendigerweise einen hohen Stellenwert haben.

3. Charakter von Evaluationssystemen

Es gibt keine Entwicklung zu einem mehr oder weniger einheitlichen Hochschulevaluationssystem, sondern eine kaum überschaubare Vielfalt. Bisher liegen keine Studien zur Evaluationspraxis an Hochschulen vor, die die Erfahrungen systematisch mehr oder weniger erschöpfend auswerten. In Übersicht 2 wird der Versuch unternommen, die wichtigsten Dimensionen zu benennen, nach denen sich die Vielfalt der vorfindlichen Hochschulevaluationen kategorisieren lassen.

So mögen die Evaluationen regelmäßig alle vergleichbaren Institutionen erfassen oder nur einen Teil von ihnen; die Teilnahme mag auch freiwillig sein. Sie mögen sich allein auf Lehre und Studium oder auf die Forschung beziehen, sie mögen multifunktional angelegt sein oder sich sogar auf die strategischen und organisatorischen Aktivitäten der Hochschule insgesamt richten. Von Fall zu Fall ist es unterschiedlich, in welchem Maße der Input, der Prozeß oder der Ertrag des Gegenstandsbereichs in Augenschein genommen wird oder auch unklare Mischebenen wie z. B. "Performanz". Die Evalution mag sich auf kleine Einheiten für Forschung und Lehre, auf Fakultäten oder auf die Hochschule insgesamt beziehen. Es kann um die Bewertung der Spitze, des Durchschnitts oder der leistungsschwachen Komponenten gehen. Man mag den Zustand des Bereichs mit Hilfe von Indikatoren, Einschätzungen von Experten, Einschätzungen der Akteure oder durch Messungen von Leistungen zu bestimmen versuchen. Man mag Mindestqualität, Rangstufungen nach Spitzenkritierien, die Verwirklichung selbstgesetzter Ziele u. a. m. mehr zu ermitteln beabsichtigen. Diese Liste der Dimensionen läßt sich ergänzen. Nicht zuletzt ist von Bedeutung, ob am Schluß des Evaluationsprozesses direkte Sanktionen erfolgen oder nicht, ob – im ersteren Falle – den besonders positiv Bewerteten weitere Privilegien zugestanden werden oder ob umgekehrt den Leistungsschwächeren Ressourcen zur Behebung ihrer Schwächen bereitgestellt werden.

Übersicht 2: Phänomene der Hochschulevaluation

Dimension		Ausprägungen
(1)	Anlaß	Diskontinuierlich negativ, diskontinuierlich, positiv, periodisch/regelmäßig
(2)	Verbreitungsgrad	Einzelfälle, generell optional, generell verpflichtend
(3)	Gegenstandsbereich	Lehre, Forschung, multifunktional, institutionell
(4)	Leistungsebene	Input, Prozeß, Output, Outcome, Impact; Effektivität, Effizienz
(5)	Institutionelle Ebene	(Person) Einheit für Lehre/Forschung, institutionell, über-institutionell
(6)	Typus des Verfahrens	Quality assurance, Quality assessment, Evaluation, Akkreditierung, (Lizensierung)
(7)	Primäre Zielebene	Spitze, Mitte, untere Regionen, Durchschnitt
(8)	Regeln der Evaluations--verfahren und -folgen	Transparent/intransparent; vorgegeben/ Gegenstand der Aushandlung; förderungs-/ selektionsbetonend; kommunikativ/nicht-kommunikativ u. a. m.
(9)	Methoden der Informationsgenerierung und Bewertung	Objektiv-subjektiv usw.
(10)	Direktheit der Messung	Indikatoren, Einschätzungen von Personen, Leistungsmessungen u. a. m.
(11)	Komplexität der Kriterien und Maße	von Einzelindikatoren bis offener Menge der Kriterien und Messungen
(12)	Bewertungsgrade und -grundlagen	Mindestqualität, Vergleich mit anderen Evaluierten (Ranking, Benchmarking, Ausmaß der Zielerreichung, "Value added" -Analyse)
(13)	Methode der Bildung einer Gesamtbewertung	Arithmetisch, konfigurativ usw.
(14)	Evaluatoren	Intern, intern/extern, extern, Wissenschaftler, "andere Experten", Vertreter der Verwendung, Staatsvertreter
(15)	Abschluß des Evaluationsverfahrens	Kommunikation zwischen Evaluatoren und Evaluierten über vorläufige Ergebnisse; Offenheit/Gerichtetheit der Beurteilung; Nicht-Öffentlichkeit/Öffentlichkeit u. a. m.
(16)	Folgen der Evaluation	Offenheit, mittel- bzw. kurzfristige Sanktion; Betonung positiver/negativer Sanktion; "Matthäus"- oder "Robin Hood"-Prinzip u. a. m.

4. Kontext der Hochschulevaluation

Angesichts der bestehenden Vielfalt von Hochschulevaluationssystemen, die wir in verschiedenen Ländern oder innerhalb einzelner Länder in verschiedenen Sektoren vorfinden, stellt sich natürlich die Frage, welche Faktoren dafür ausschlaggebend sind, daß bestimmte Evaluationssysteme realisiert werden sollen oder sich tatsächlich durchsetzen. Auch hier gibt es bisher nur wenige Analysen; Forschung über die Entwicklung von Evaluation in Hochschule und Forschung hat ein breites Betätigungsfeld vor sich.

Vieles spricht dafür, daß sich Hochschulevaluationssysteme grundlegend danach unterscheiden,

- welche vor-evaluativen Bewertungen bei der Einführung der Evaluation über den Zustand der zu evaluierenden Bereiche herrschten und in welchem Maße Veränderungen für notwendig gehalten wurden,

- in welchem Maße eine substantielle Einheitlichkeit oder substantielle Vielfalt im Hochschulwesen bestand und ob eine weitgehende Einheitlichkeit der Qualität oder große Qualitätsdifferenzen zu beobachten waren,

- inwieweit sich eine Evaluationskultur entwickelt und entfaltet hat und

- welche nationalen kulturellen Voraussetzungen für Evaluation bestehen.

Für die letztgenannte Dimension unternahm der bekannte Evaluationsspezialist Herb Kells den Versuch einer Klassifikation in Anlehnung an Hofstede. Wie Übersicht 3 zeigt, scheinen das Ausmaß der in einer Gesellschaft vorherrschenden Machtdistanz und der Unsicherheitsvermeidung sowie die Stellung der Kultur zwischen "Maskulinität" und "Feminität" sehr bedeutsam zu sein.

In Hochschulsystemen, die sich in Frühphasen der Etablierung einer Evaluationskultur befinden, scheint von großer Bedeutung zu sein, welche Absichten zur Gestaltung des Hochschulsystems mit dem Auf- und Ausbau von Evaluierungsaktivitäten verfolgt werden:

- Wird der bestehende Kontext, wird die bestehende "Logik" des Hochschulsystems weitgehend akzeptiert und in diesem Rahmen eine höhere Reflexivität und Lernfähigkeit des Hochschulsystems angestrebt?

- Wird zwar vermutet, daß eine solche Reflexivität und gesteigerte Lernfähigkeit zu größeren Veränderungen des Hochschulsystems führen kann, aber die Richtung solcher Veränderung diesem Lernprozeß selbst überlassen?

- Wird das bestehende Hochschulsystem in seinen Grundzügen akzeptiert, aber mit der Einführung bzw. dem Ausbau von Evaluationssystemen das Ziel verfolgt, im bestehenden System stärkere Prioritäten zu setzen?

- Wird die Etablierung eines Evaluationssystems als Instrument verstanden, das Hochschulsystem in die Richtung der eigenen Zielvorstellungen zu lenken? Wird also Evaluation mit system-interventionistischen Zielsetzungen etabliert?

Sehr oft können wir beobachten, daß die Einführung von Evaluationssystemen auf der Basis von stark system-interventionistischen Vorstellungen betrieben wird. Dies löst oft bei einem großen Teil der Beteiligten ein so großes Mißtrauen aus, daß die Entfaltung einer Evaluationskultur darunter leid Sehr oft zeigt sich dabei jedoch, daß zwar bestimmte System-Interventionen intendiert sind, jedoch Instrumentarien der Evaluation gewählt werden, die den eigenen Zielen nicht entsprechen oder in dem bestehenden Kontext nicht den erwarteten Erfolg haben können. Auf der Basis des Vergleichs einer großen Zahl von nationalen Hochschulevaluationssystemen kam Kells (1999) zu dem Schluß, daß die Mehrheit der bestehenden Systeme weder in den nationalen Kontext passen noch hinreichend an den gesetzten Zielen ausgerichtet sind.

Vielleicht läßt sich die These vertreten, daß Evaluationssysteme in ihren Frühphasen zumeist kontext-erratisch sind. Im Laufe der Zeit werden dann nicht nur anfängliche Fehler erkannt, sondern die Ambitionen, mit Hilfe von Evaluationsmechanismen die "Logik" der Hochschulsysteme zu verändern, werden bescheidener.

Übersicht 3: Dimension von evaluationsrelevanten Kulturen
(in Anlehnung an Hofstede)

High Power Distance	High Avoidance of Uncertainty	Masculinity
• heavy use of objective measures, performance indicators • use of ranking and comparison • heavy accountability • public reporting • system designed by people in power, centrally • peer review highly important • powerful offices and figures rarely evaluated	• standards and firm even prescriptive, objective criteria • use of ranking and comparison • favor the best • tendency towards one high standard, one model • deadlines and firm schedules for review • take long time to design system • short perspective	• money is important; link eval. to funds where possible • use sanctions; make choices • favor the best • use of performance measures • focus on rules, schedules and compliance • self-eval is more difficult; public acknowledgement of dysfunction rare
Low Power Distance	**Low Avoidance of Uncertainty**	**Femininity**
• many professionals help design the system • all levels and parts are included • self-eval, internal focus • all get evaluated • improvement is the focus • national system best based on university-based culture and internal eval. Systems	• long term perspective cycles used • learn-by-doing; 'get started'; protection • low reliance on standards • high focus on achievement of stated intentions and on use of guild expertise	• few if any sanctions; some incentives used • all are helped; 'best' are *not* favored • little public reporting • training provided • change to come through 'discovery' and acquired 'psych. ownership'

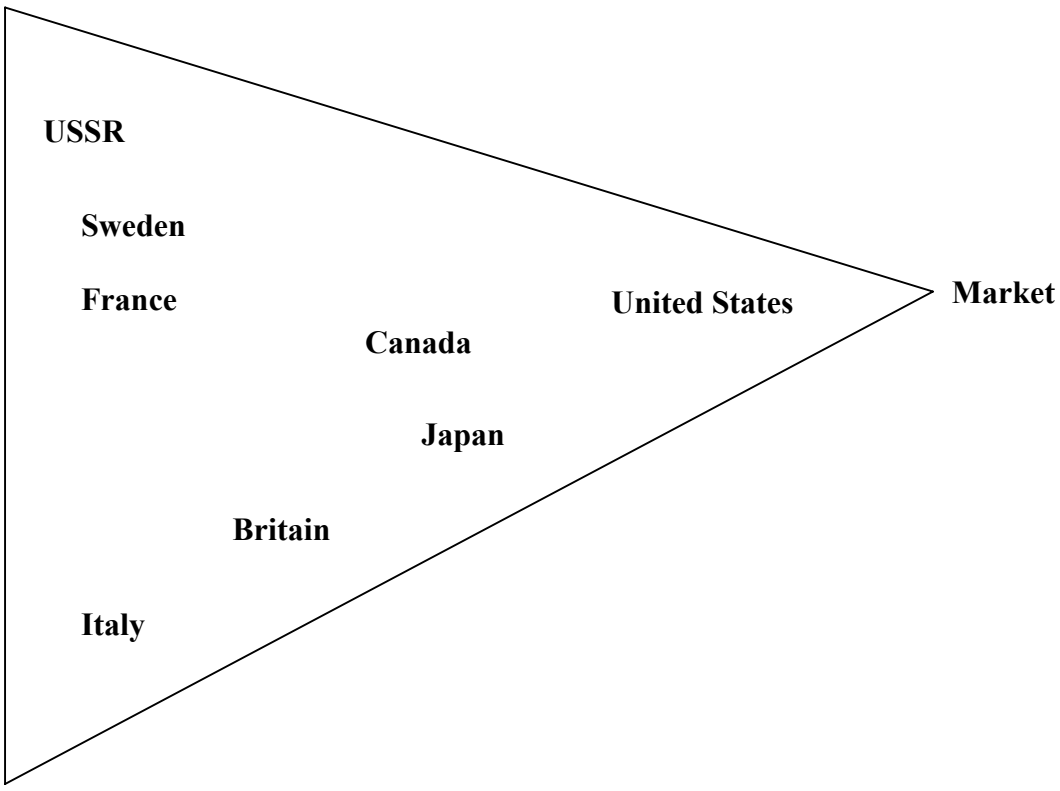
Quelle: Kells (1999)

5. Evaluation und Veränderung der Machtverhältnisse

Die Ausbreitung von Hochschulevaluation ging einher mit einer wachsenden Komplexität der Machtverhältnisse. So nannte Burton Clark Anfang der achtziger Jahre drei Mächte, die das Hochschulwesen prägen (siehe Übersicht 4): die Autorität des Staates, der Markt und die „academic oligarchy".

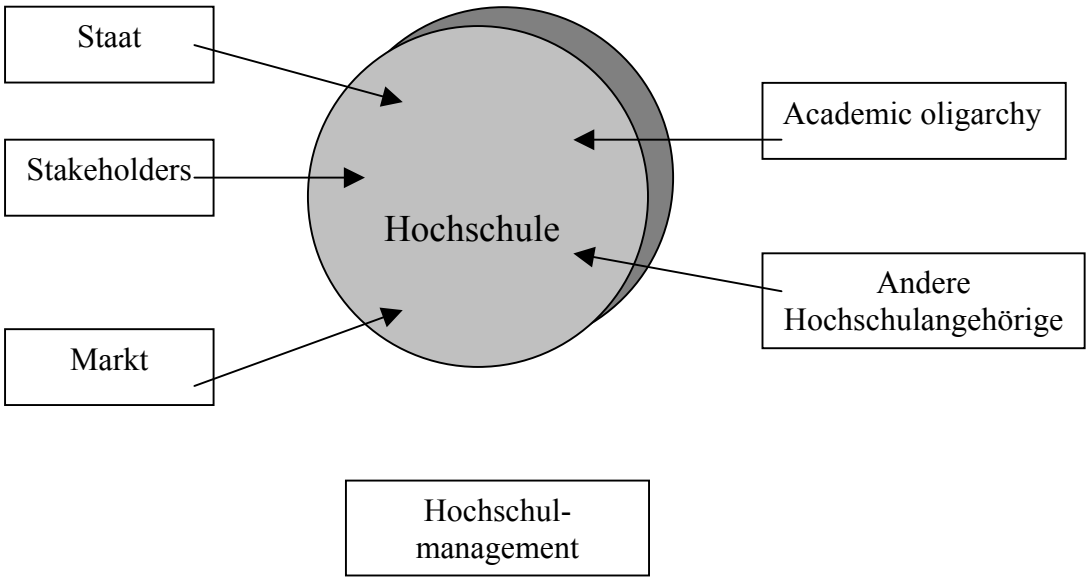
Heute erscheint es eher angemessen, von sechs Mächten zu sprechen (siehe Übersicht 5): Hinzuzuzählen sind die „stakeholders", andere Hochschulangehörige und, last not least, das Hochschulmanagement.

Übersicht 4: "The Triangle of Coordination"



Quelle: Clark (1983).

Übersicht 5: Akteure und Mächte



Quelle: Teichler (1998).

Die Zunahme der Mächte kompliziert verstärkt sowohl die Evaluation als auch Entscheidungen auf der Basis von Evaluationsergebnissen. Um im Zuge wachsender Machtkomplexität überhaupt eine Handlungsfähigkeit zu sichern, wurden immer komplexere Regelungssysteme zur Gestaltung der Handlungsmöglichkeiten und Machtkonfigurationen etabliert. Diese lassen sich beschreiben nach

- Akteuren (z. B. Stärkung des Hochschulmanagements, Schwächung der Academic oligarchy),

- institutionellen Ebenen (z. B. Stärkung der institutionellen Ebene in Kontinental-Europa zu Lasten der staatlichen und Lehr-/Forschungseinheits-Ebene; Stärkung der überinstitutionellen Ebene in GB und USA),

- Prozessen (z. B. das niederländische Steuerungsmodell der achtziger Jahre: stärkere Zielvorgabe, geringere Prozeßkontrolle, größere Rolle der Evaluation).

In den meisten Ländern scheint parallel zum Auf- und Ausbau von Hochschulevaluation eine Stärkung des Hochschulmanagements betrieben zu werden – gleichgültig, ob den Hochschulleitungen eine wachsende Autonomie zur institutionellen Profilbildung eingeräumt wird oder sie nur mehr institutionelle Handlungsspielräume unter starken externen Vorgaben erhalten. Das Hochschulmanagement ist der Hoffnungsträger, daß die Reflexion der Stärken und Schwächen zunimmt und Aktion zur Folge hat.

6. Die "Mission impossible" des Hochschulmanagements

Vom Hochschulmanagement werden eine Fülle zentraler Leistungen erwartet, um Hochschulevaluation im Hinblick auf Forschung und Lehre funktionsfähig zu machen. Vor allem folgende Erwartungen werden in aktuellen Debatten immer wieder sichtbar:

- Das Hochschulmanagement hat als Gelenkstelle zwischen dem Innenleben der Wissenschaft und der Gesellschaft den Hochschulangehörigen die offene Selbstkritik der Evaluation schmackhaft zu machen und der Außenwelt die großen interventionistischen Hoffnungen, die sie mit der Evaluation verbinden, auszureden, damit Evaluation machbar wird.

- Das Hochschulmanagement hat eine Balance zwischen Vielfalt, Dezentralität und systemischen Grenzen valider Evaluation von Innovation und Qualität einerseits und Wunsch nach institutionell konsistentem Profil andererseits anzustreben.

- Das Hochschulmanagement hat dazu beizutragen, daß die gewöhnlich bestehende Inbalance in der Leistungsbewertung verschiedener Funktionen der Hochschule sich nicht in einer Inbalance der Leistungen niederschlägt.

Manches spricht dafür, daß das Hochschulmanagement in der Regel mit solchen Erwartungen überfordert ist. Die Fachliteratur zur Rolle des Universitätspräsidenten zeichnet deren Ambivalenz zwischen "Chamäleon" und "Herkules" eindrucksvoll nach. Auch wird in den letzten Jahren immer stärker betont, daß zu große Heilserwartungen auf einen "Managerialismus" der Hochschulen gesetzt worden sind, die den Flexibilitätserfordernissen einer dezentralen professionellen Organisation nicht entsprechen und selbst für das Management von kommerziellen Produktions- und Dienstleistungsoganisationen als überholt erscheinen.

Hinzu kommt, daß das Hochschulmanagement mit asymmetrischen Gestaltungspotentialen gegenüber den Hauptfunktionen der Hochschule ausgestattet ist: mit weitaus größeren Gestaltungspotentialen gegenüber der Lehr- als der Forschungsfunktion. Dazu tragen verschiedene Faktoren bei:

- die höhere Priorität der Lehre in Hochschulpolitik und -planung,

- das fehlende Monopol bzw. Oligopol der Hochschulen in der Forschung,

- die unvollständige interne Finanzierung der Forschung, die eine starke externe Orientierung der Hochschulangehörigen zur Folge hat,

- starke Macht überinstitutioneller Kriteriensetzung in der Beurteilung der Forschung.

Diese Unvollständigkeit der Gestaltungspotentiale ist in Deutschland relativ ausgeprägt. Dazu tragen die Verfassungsverankerung der Forschungsfreiheit, die Tradition des Lehrstuhls sowie die Etablierung eines starken Sektors der Grundlagenforschung (so insbesondere die Max-Planck-Gesellschaft) außerhalb der Hochschulen bei.

Damit gibt sich für das Hochschulmanagement zunächst – auf den ersten Blick – ein Evaluations- und Entscheidungsparadox. Die Forschung, die sich im Prinzip eher eindeutigen Evaluations- und Entscheidungskriterien unterziehen läßt, ist dem Hochschulmanagement in der Entscheidung stärker entrückt als die Lehre, die weniger eindeutig evaluier- und entscheidbar ist.

Für das Hochschulmanagement in Deutschland ergibt sich zweifellos die Aufgaben, Lehr- und Lernevaluation in den Mittelpunkt zu rücken, damit in diesem Gebiet überhaupt eine regelmäßige und systematische Reflexion der Situation erfolgt und auch mit sanften Anreizen und Sanktionen verbunden wird, damit nicht Bemühungen um Qualität der Lehre immer gegenüber Bemühungen um Qualität der Forschung an den Rand gedrängt werden. Forschungsevaluation wird unvermeidlich vielfältig, heterogen und zumindest zum Teil fremdbestimmt bleiben: Das Hochschulmanagement kann hier soweit Akzente setzen, daß es die Legitimität seiner Bemühungen um Profilbildung der Hochschulen steigert.

Literaturverzeichnis

Bargh, C., Scott, P. and Smith, D. (1996), Governing Universities. Changing the Culture? Buckingham

Brinckmann, H. (1998), Die neue Freiheit der Universität. Operative Autonomie für Lehre und Forschung an Hochschulen, Berlin

Clark, B. R. (1983), The Higher Education System. Academic Organization in Cross-National Perspective, Berkeley und Los Angeles, CA

Clark, B. R. (1998), The Entrepreneural University, Oxford

Daniel, H.-D. (1998), Beiträge der empirischen Hochschulforschung zur Evaluierung von Forschung und Lehre, in: Teichler, U., Daniel, H.-D. und Enders, J.: Brennpunkt Hochschule, Frankfurt a.M. und New York, S. 11-53

In't Veld, R., Füssel, H.-P. und Neave, G. (Hg.) (1996), Relations between State and Higher Education, The Hague

Kells, H. R. (1999), National Higher Education Evaluation Systems: Methods for Analysis and Some Propositions for the Research and Policy Void, in: Higher Education, 39. Jg., H. 3 (im Druck)

Sanyal, B. C. (Hg.) (1996), Institutional Management in Higher Education, Paris, UNESCO, IIEP

Teichler, U. (1998), Managementreformen an deutschen Hochschulen. Einige Betrachtungen aus der Distanz, in: Ermert, K. (Hg.): Hochschulmanagement, Rehburg-Loccum, Loccumer Protokolle Nr. 25/98, S. 9-33

The Evaluative State Revisited (Themenheft) (1998), in: European Journal of Education, 33. Jg., H. 3

Westdeutsche Rektorenkonferenz (1989), Staatliche Steuerung und Erneuerung des Studiums, Bonn (Dokumente zur Hochschulreform, Bd. 67)

Programm des Workshops

Donnerstag, 20. Mai 1999

15:30 Uhr **Eröffnung und Einführung**

Meinolf Dierkes (WZB)

16:00 Uhr **Evaluation im wissenschaftspolitischen Kontext**
- Ein Vergleich europäischer Aktivitäten -

Ulrike Felt (Universität Wien)

17:30 Uhr **Unterschiedliche Aufgaben – gemeinsame Ziele?**
- Entwicklung und Bewertung der WGL-Institute -

Ekkehard Nuissl von Rein
(Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz – WGL)

19:00 Uhr Buffet

Freitag, 21. Mai 1999

9:00 Uhr **Input, Throughput, Output:**
Welche Indikatoren zu welchem Zweck?

Martina Röbbecke (WZB)

Stefan Hornbostel
(Centrum für Hochschulentwicklung – CHE)

11:00 Uhr **Verfahren der Selbstevaluation von Forschung**

Dagmar Simon (WZB)

Jürgen Lüthje (Universität Hamburg)

Adrian Verkleij
(Center for Higher Education Policy Studies – CHEPS)

13:00 Uhr Mittagspause

14:00 Uhr **Forschungsmanagement im Wandel**

Ulrich Teichler
*(*Wissenschaftliches Zentrum für Berufs- und Hochschulforschung, Universität/Gesamthochschule Kassel)

15:30 Uhr Kaffee-/Teepause

16:00 Uhr **Auswertung und künftige Perspektiven**

Diskussionsleitung:
Helmut Wollmann (Humboldt-Universität)

Teilnehmerinnen und Teilnehmer

Juliane Andersohn	Forschungsverbund Berlin e.V. Rudower Chaussee 5 12524 Berlin
Prof. Dr. Hans-Dieter Daniel	Wissenschaftliches Zentrum für Berufs- und Hochschulforschung Universität Gesamthochschule Kassel Henschelstr. 4 34109 Kassel
Prof. Dr. Meinolf Dierkes	Direktor der Abteilung "Organisation und Technikgenese", WZB
Dr. Bernd Ebersold	Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V. Hofgartenstr. 2 80539 München
Dr. Jürgen Enders	Wissenschaftliches Zentrum für Berufs- und Hochschulforschung Universität Gesamthochschule Kassel Henschelstr. 4 34109 Kassel
Prof. Dr. Ulrike Felt	Institut für Wissenschaftstheorie und Wissenschaftsforschung Universität Wien Sensengasse 8 / 10 A-1090 Wien
Dr. Dirk Hartung	Max Planck Institut für Bildungsforschung Lentzeallee 94 14195 Berlin
Dr. Wieland Hempel	Senatsverwaltung für Wissenschaft, Forschung und Kultur Brunnenstr. 188/190 10119 Berlin
Dr. Stefan Hornbostel	CHE Centrum für Hochschulentwicklung Postfach 105 33311 Gütersloh

Dr. Axel Horstmann	Volkswagen-Stiftung Kastanienallee 35 30519 Hannover
Dr. Erwin Jost	Max-Delbrück-Centrum für Molekulare Medizin (MDC) Berlin-Buch Robert-Rössle-Str. 10 13092 Berlin
Dr. Jörg Klein	Geschäftsstelle der Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz Ahrstr. 45 53175 Bonn
Klaus Lömker	Bundesministerium für Bildung und Forschung Referat 321 Heinemannstr. 2 53175 Bonn
Dr. Dr. h.c. Jürgen Lüthje	Präsident der Universität Hamburg Edmund-Siemers-Allee 1 20146 Hamburg
Sandra Mittag	Verbund Norddeutscher Universitäten Universität Hamburg Edmund-Siemers-Allee 1 20146 Hamburg
Prof. Dr. Ekkehard Nuissl von Rein	Wissenschaftlicher Vizepräsident der WGL Deutsches Institut für Erwachsenenbildung Hansaallee 150 60320 Frankfurt a.M.
Dr. Maria Oppen	WZB, Abteilung "Regulierung von Arbeit"
Dr. Martina Röbbecke	WZB, beim Präsidenten
Dr. Dagmar Simon	WZB, Forschungsplanung und -koordination
Barbara M.-L. Steiger	Hochschulrektorenkonferenz Ahrstraße 39 53175 Bonn
Dr. Andreas Stucke	Geschäftsstelle des Wissenschaftsrates Brohler Str. 11 50968 Köln

Prof. Dr. Ulrich Teichler	Wissenschaftliches Zentrum für Berufs- und Hochschulforschung Universität Gesamthochschule Kassel Henschelstr. 4 34109 Kassel
Dr. Georg Thurn	WZB, Forschungsplanung und -koordination
Adrian Verkleij	Center for Higher Education Policy Studies University of Twente (CHEPS) P.O. Box 217 7500 AE Enschede Niederlande
Dr. Beatrix Vierkorn-Rudolph	Geschäftsstelle der Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz Ahrstr. 45 53175 Bonn
Heinz Westkamp	Bundesministerium für Bildung und Forschung Referat 424 Heinemannstr. 2 53175 Bonn
Dr. Ekkehard Winter	Stifterverband für die Deutsche Wissenschaft Barkhovenallee 1 45239 Essen
Prof. Dr. Hellmut Wollmann	Philosophische Fakultät III Humboldt-Universität zu Berlin Clara-Zetkin-Str. 28 10099 Berlin