In: Jürgen Kaube (Hrsg.), Die Illusion der Exzellenz. Lebenslügen der Wissenschaftspolitik. Berlin: Wagenbach (2009): 65-73.

Das *Peer Review*-System auf dem ökonomischen Prüfstand

Das Wissenschaftssystem in Deutschland, Österreich und der Schweiz ist zu einem hohen Ausmaß von *Peer Reviews* abhängig geworden, von der gegenseitigen Evaluation wissenschaftlicher Leistungen durch Kollegen. Die Karriere, insbesondere die Berufung zum Professor oder zur Professorin, sowie die Einwerbung von Drittmitteln hängt entscheidend vom Abschneiden im *Peer Review*-Prozess ab. Neuerdings wird auch die Besoldung von der Publikationsleistung abhängig gemacht. Inzwischen hat das *Ranking*- und *Rating*-System eine besondere Eigendynamik entfaltet, welche die Wissenschaft zunehmend einer externen Kontrolle unterwirft, die ihrerseits zumeist auf durch *Peers* begutachtete Publikationen beruht.

Gegenüber dem bisherigen Hochschulsystem stellen diese Entwicklungen Reformen dar, die durch leistungsorientierte Anreizsysteme gekennzeichnet sind. Universitäten, heißt es, sollen unternehmerischer werden. Will man die Erfolgschancen dieser Reformen einschätzen, sollte man die Unterschiede zwischen der Wissensproduktion in gewinnorientierten Unternehmen und dem durch öffentliche Gelder geförderten Wissenschaftssystem in Rechnung stellen.

Das Belohnungs- und Anreizsystem der Forschung

Die Funktionalität eines wissenschaftlichen Belohnungs- und Anreizsystems zeichnet sich durch vier Merkmale aus:

(1) Ein funktionierendes *Peer Review*-System hat überragende Bedeutung für die Entwicklung und Beurteilung der Qualität von Wissenschaft und damit für das Belohnungs- und Anreizsystem. Nur Wissenschaftler können die Qualität von Wissenschaft beurteilen.

(2) In der Wissenschaft sollte der fixe Anteil des Einkommens hoch sein. Der Grund liegt in dem großen Risiko, nicht der oder die Erste bei einer Entdeckung zu sein. Die jeweils Zweiten können einen empfindlichen Verlust erleiden. Innerhalb von Forschungsteams sollten zudem hohe Lohndifferenzen vermieden werden, um die Kooperationsbereitschaft aufrechtzuerhalten. Empirische Befunde zeigen denn auch, dass hohe Lohndifferenzen die Teamproduktivität mindern. Dies gilt insbesondere in der Forschung.

(3) In der Wissenschaft spielt die nichtmonetäre Belohnung eine große Rolle. Es gibt einen »taste for science« (Robert K. Merton). Zwei nichtmonetäre Anreize sind besonders bedeutsam: Die Anerkennung durch die wissenschaftliche Gemeinschaft in Form von Preisen oder Ehrendoktoraten; für die meisten Wissenschaftler sind allerdings Publikationen und Zitationen die relevante Form der Anerkennung. Zum anderen ist die gewährte Autonomie ein zentraler Anreiz. Sie ist die wichtigste Voraussetzung für kreative Arbeit und hohen Forschungsoutput. Forscher nehmen Einkommenseinbußen in Kauf, wenn sie ein größeres Maß an Autonomie erhalten.

(4) Das im Vergleich zu anderen Arbeitsmärkten niedrigere Primäreinkommen an der Hochschule hat einen Selektionseffekt. Es zieht diejenigen an, die eine hohe Präferenz für Autonomie und Anerkennung durch die *Peers* haben – beides Voraussetzungen für eine erfolgversprechende wissenschaftliche Tätigkeit.

Das Peer Review-System auf dem Prüfstand

Das traditionelle Begutachtungssystem steht schon seit geraumer Zeit aufgrund umfangreicher empirischer Evidenz auf dem Prüfstand. Die Kritik lässt sich folgendermaßen zusammenfassen: Die Übereinstimmung zwischen Gutachterurteilen ist gering. Bleiben die ablehnenden Gutachten unberücksichtigt, ist der Grad an Seriosität bei der Auswahl der Besten sogar minimal. Die Gutachten haben zudem nur eine geringe prognostische Kraft: Das Urteil, das die Gutachter über die Qualität eines Manuskriptes abgeben, korreliert nur äußerst schwach mit den späteren Zitationen. Außerdem beurteilen Gutachter Artikel besser, die ihre eigenen Arbeiten zustimmend zitieren. Entsprechend fühlen sich Autoren oft von Gutachtern gedrängt, ihre Manuskripte zu ändern, um die Akzeptanz bei den Gutachtern zu erhöhen, auch wenn diese Änderungen ihrer Überzeugung widersprechen. Nach einer Untersuchung von Bedeian gilt dies bei nicht weniger als 25 Prozent der Autoren.[1] Man kann ein solches Verhalten »akademische Prostitution« nennen. Das System der qualitativen *Peer Reviews* beruht mithin auf einer erstaunlich fragwürdigen wissenschaftlichen Grundlage.

Als Reaktion auf die Kritik an qualitativen *Peer Reviews* haben sich zunehmend quantitative, bibliometrische Verfahren etabliert, die auf Zitationsanalysen und *Impact*-Faktoren beruhen. Von quantitativen Methoden erhofft man sich zum einen, dass sie objektiver als qualitative Methoden sind, weil sie, so die Annahme, auf weitaus mehr Einschätzungen beruhen als die drei bis vier üblichen *Peer Reviews*, weil sie Begünstigungen durch *old boys' networks* vermeiden beziehungsweise kontrollieren; und weil sie nicht-reaktiv sind, das heißt keine Rückwirkungen auf den *Review*-Prozess haben. Zum anderen wird als Vorteil quantitativer Messgrößen angesehen, dass sie auch von Personen außerhalb der *Republic of Science* herangezogen werden können, insbesondere von fachfremden Kollegen, der Presse, der Wissenschaftsadministration und der Politik. Das erklärt die Beliebtheit bibliometrischer Verfahren als Grundlage von *Rankings* und *Ratings* und erklärt auch den Druck der Politik, *Rankings* beziehungsweise *Ratings* als Ausgleich für die größere Autonomie zu etablieren, die den Hochschulen neuerdings gewährt wird.

Allerdings sind die genannten Vorteile quantitativer, bibliometrischer Verfahren durchaus problematisch. Eine vorläufige Bestandsaufnahme der Probleme bibliometrischer Verfahren in der Literatur lässt drei Gruppen solcher Probleme unterscheiden, die kumulativ wirken.

Zum Ersten bilden dieselben *Peer Reviews*, deren Problematik überwunden werden soll, die Basis aller bibliometrischen Verfahren. Denn gezählt werden zumeist nur Veröffentlichungen in begutachteten *Journals*.

Zum Zweiten gibt es beträchtliche methodische Probleme.

(1) *Selektionsprobleme:* Bibliometrische Daten repräsentieren immer nur einen Teil des wissenschaftlichen Kommunikationsprozesses. Die Dominanz der Indizes hat dazu geführt, dass die Selektionskriterien, die Datenqualität und die Methoden der Datenaufbereitung dieser Indizes kaum mehr kritisch in die Einschätzung der Ergebnisse einbezogen werden.

(2) *Mangelndes Wissen über Zitier-Gewohnheiten:* Diese Gewohnheiten sind nicht nur in einzelnen Disziplinen, sondern auch in Subdisziplinen sehr unterschiedlich, sodass ein Vergleich bibliometrischer Daten schon innerhalb ein und derselben Disziplin problematisch sein kann. Darüber hinaus hängt die Zahl der Zitationen von der Erreichbarkeit der Quellen ab.

(3) *Mangelndes Wissen über die Art der Zitate:* Diese können eine zustimmende oder ablehnende Bedeutung haben. Mitunter folgt die Zitierweise einfach einem Herdenverhalten oder spiegelt die akademische Popularität von Star-Papieren wider. Viele Zitate werden nur deshalb eingefügt, um Gutachter freundlich zu stimmen. Simkin und Roychowdhury zeigen anhand einer Analyse übertragener Fehler, dass etwa 70 bis 90 Prozent aller zitierten Papiere vermutlich gar nicht gelesen wurden.[2]

(4) *Irreführung durch den Impact-Faktor:* Die relative Bedeutung einer wissenschaftlichen Zeitschrift, gemessen an der Häufigkeit, mit der diese Zeitschrift andernorts zitiert wird, besagt nicht, dass die in ihr veröffentlichten Artikel diesem Durchschnitt entsprechen. Adler, Ewing und Taylor stellen denn auch in einem Gutachten für die *International Mathematical Union* fest, dass die Verwendung von *Impact*-Faktoren zu großen Fehlerwahrscheinlichkeiten führen könne und »atemberaubend naiv« sei.[3] Angesichts dieser methodischen Probleme erstaunt die geringe Übereinstimmung der verschiedenen *Rankings* nicht. Für die Ökonomik haben Ursprung und Zimmer gezeigt, dass im einflussreichen *Ranking* deutscher akademischer Ökonomen, die das »Handelsblatt« vornimmt, jede dritte der einbezogenen Personen an die Spitze vorrücken könnte, wenn man die Methode ändert.[4]

Die dritte Problemgruppe ist zugleich diejenige, die am wenigsten diskutiert wird. Bibliometrische Verfahren geben Wissenschaftlern und wissenschaftlichen Institutionen einen starken Anreiz, sich strategisch zu verhalten. Es soll hier nur ein

Beispiel genannt werden: Hochschulen in Großbritannien und den USA engagieren hochbezahlte Forscher-Stars, insbesondere kurz vor Evaluationen, um in *Rankings* und *Ratings* gut abzuschneiden. Die dafür aufzuwendenden Mittel werden eingespart, indem weniger Nachwuchskräfte eingestellt und gefördert werden und die Anzahl der *Postdocs* je Vollprofessur wächst.

Quantitative Evaluationskriterien bergen immer die Gefahr in sich, dass infolge ihrer Nutzung professionelle ethische Normen erodieren. Hochgradig verzerrend ist auch der *multiple tasking effect*: Komplexe Aufgaben sind durch eine große Anzahl verschiedener Kriterien gekennzeichnet, die sich unterschiedlich gut quantifizieren lassen. In der Folge werden sich die Evaluierten in erster Linie an den leicht messbaren Kriterien orientieren und die schwer messbaren außer Acht lassen, obwohl diese häufig die wichtigeren sind.

Im Hochschulbereich bedeutet dies, dass vielfach Qualität und Innovativität auf dem Altar der Quantität von Publikationen geopfert oder dass Forschungsmittel verschwendet werden. Forscher wenden die »Salamitaktik« an, indem sie neue Ideen oder interessante Datensätze so dünn wie Salamischeiben aufschneiden und anbieten, um die Anzahl der Publikationen zu maximieren. Dieses Verhalten wird verstärkt, wenn monetäre Belohnungen an die Zahl von Publikationen geknüpft werden. Konventionelle oder modische Ansätze werden vorgezogen, weil sie wenig Widerspruch hervorrufen und leichter zu veröffentlichen sind. Dies führt zu einer Homogenisierung der gesamten Forschung, wie dies bereits für *Business Schools*[5] und die Volkswirtschaftslehre[6] nachgewiesen wurde. Die Orientierung an *Impact*-Faktoren bewirkt überdies, dass kleinere, spezialisierte und nicht-englischsprachige Zeitschriften an Attraktivität verlieren. Wird die Veröffentlichung in Zeitschriften mit hohem *Impact*-Faktor als Kriterium für die Beurteilung von Fachbereichen herangezogen, werden letztendlich weniger Wissenschaftler eingestellt, die Randbereiche vertreten und in solchen kleineren Spezialzeitschriften publizieren. Schließlich werden nur noch erfolgreiche Tests publiziert und negative Ergebnisse verschwiegen oder gar beseitigt, weil sie sich schlechter publizieren lassen. Dies widerspricht dem Ideal einer Wissenschaft, welche die Falsifikation von Hypothesen als ihre Kernaufgabe an-

sieht. Lernen aus den Fehlern der Forschungsgemeinschaft ist nicht mehr möglich. Eine weitere, indirekte Folge der Orientierung an Quantität anstelle schwer messbarer Qualität ist die wachsende Last der Begutachtung und der damit einhergehenden Wahrscheinlichkeit, dass die Qualität der Gutachten sinkt. Gutachten werden immer häufiger an weniger qualifizierte Forscher weitergereicht, die nicht in der Lage sind, die Innovativität einer Arbeit richtig einzuschätzen.

Vielen Forschern und Forschungsinstitutionen wie zum Beispiel dem deutschen Wissenschaftsrat gelten darum *informed Peer Reviews* als Königsweg, die Kombination also aus quantitativen und qualitativen Evaluationsverfahren. Bibliometrische Verfahren sollen qualitative *Peer Reviews* korrigieren und ergänzen. Jedoch ist es keineswegs sicher, dass eine Kombination quantitativer und qualitativer Kriterien zum Beheben der Fehlerquellen führt. Ebenso gut ist möglich, dass diese Kombination bei fahrlässigem Gebrauch der Kriterien kumulativ oder sogar multiplikativ wirkt. Die Indikatoren haben dann eine beachtliche, nicht kontrollierte Hebelwirkung.

Wirkungen einer Verstärkung monetärer Anreize

Die deutschsprachigen Länder sind im Begriff, das *pay for performance* in das Hochschulsystem einzuführen, beispielsweise durch die sogenannte W-Besoldung in Deutschland. Aus den geschilderten Besonderheiten der Wissensproduktion in der Forschung ergeben sich jedoch vier schwerwiegende Einwände gegen diesen Schritt.

(1) *Die Ungleichheit zwischen den Einkommen wächst:* Diese Ungleichheit bewirkt, dass die für wissenschaftliche Arbeit zunehmend wichtiger werdende Teamproduktion gefährdet wird. Auch das geschilderte opportunistische Verhalten von Gutachtern und Autoren dürfte Auftrieb erhalten.

(2) *Der risikobehaftete Anteil des Einkommens wächst:* Unterstellt wird, dass monetäre Anreize Anlass geben, die Anstrengung in die erwünschte Richtung zu verstärken. Diese Annahme hat sich schon im Bereich der Management-Entlohnung als äußerst fragwürdig erwiesen.

(3) *Der Anreiz zu strategischem Verhalten wächst:* Wird das monetäre Interesse von Forschern gestärkt, dann sollte damit gerechnet werden, dass ökonomisch denkende Forscher ihr Einkommen mit jenen Mitteln aufzubessern suchen, die am wenigsten Aufwand erfordern, also gerade nicht durch Investition in innovative Forschung. Der Anreiz nimmt zu, das System der Erfolgszuschreibung für sich zu instrumentalisieren. Dadurch entsteht ein *Lock-in*-Effekt, der es auch Kritikern dieses Systems immer schwerer macht, sich ihm zu entziehen.

(4) *Intrinsische Motivation wird verdrängt:* Monetäre Anreize können unter bestimmten Bedingungen die für die Forschung entscheidende intrinsische Motivation verdrängen. Außerdem kann auch eine verstärkte extrinsische Motivation den Verlust an vormals gegebener intrinsischer Motivation nicht ausgleichen. Dieser Effekt führt zugleich zu einer Verringerung der empfundenen Autonomie. Intrinsische Motivation kann deshalb in erster Linie durch Erhöhung der Autonomie gesteigert werden. Eine »Leistungsentlohnung« hingegen entzieht den Wissenschaftlern das Vertrauen, eigenverantwortlich eine große Leistung zu erbringen. Diese Misstrauenskundgebung verringert die Loyalität gegenüber der beschäftigenden Institution.

Ein radikaler Vorschlag

Angesichts der schwerwiegenden Mängel des *Peer Review*-Systems stellen wir einen radikalen Vorschlag zur Diskussion. Er besteht aus drei Komponenten:

(1) Die Rücknahme von *pay for performance* im Wissenschaftssystem: Ein hoher variabler Anteil des Einkommens widerspricht den Besonderheiten der Wissensproduktion in der Forschung. Er erhöht das ohnehin hohe Risiko des Misserfolges, verringert die Bereitschaft zur Wissensweitergabe im Team, steigert den Anreiz, die Schwächen des *Peer Review*-Systems strategisch auszunutzen, verdrängt die intrinsische Motivation und unterminiert insgesamt den Forschungsdrang. Hingegen sind die positiven Anreizwirkungen der W-Besoldung äußerst ungewiss.

(2) Die Eindämmung der Abhängigkeit vom *Peer Review*-System: Im derzeitigen System sind Forscherkarrieren total und

lebenslang von einem fehlbaren *Review*-System abhängig. Diese Abhängigkeit wird durch den »Evaluations-Hype« und durch steigende Drittmittelanteile an den Forschungsressourcen noch gesteigert. Der Einfluss der *Peer Reviews* lässt sich eindämmen, wenn die Forschenden und Lehrenden sorgfältig sozialisiert und ausgelesen werden und ihnen anschließend ein hoher Grad an Autonomie gewährt wird. Bei dieser zukunftsorientierten Strategie müssen durchaus qualitative und quantitative *Peer Review*-Verfahren zum Einsatz kommen. Dies sichert bei aller Fehlbarkeit des *Peer Review*-Systems, dass Standards der Wissenschaftlichkeit erfüllt sind, und gibt Hinweise auf das Potential der Kandidaten. Im Anschluss an die Ernennung zum Professor, die nach strengen Kriterien zu erfolgen hat, muss aber darauf vertraut werden, dass die berufene Person die erwarteten Leistungen auch ohne ständige Kontrolle erbringt. Dadurch wird die für die Wissenschaft unerlässliche Autonomie zumindest ab der Berufung auf eine volle Professur gewährleistet. Deshalb sind Berufungsverfahren das mit Abstand wichtigste Geschäft einer wissenschaftlichen Institution. Dabei ist durchaus mit einer gewissen Varianz zu rechnen. Manche der Ausgewählten werden in ihrer Leistung nachlassen, andere hingegen werden durch den gebotenen Freiraum beflügelt und zu Spitzenleistungen motiviert. Das Prinzip der strengen Auslese und der anschließenden Gewährung von Autonomie ist in den »Leuchttürmen der Wissenschaft«, wie beispielsweise der Harvard-Universität, selbstverständlich.

Forscher durchlaufen eine ungewöhnlich lange Selektions- und Sozialisationsphase, in der ihre Autonomie eingeschränkt ist. Auch nach deren Abschluss müssen sie sich bei jeder beabsichtigten Publikation und bei jedem Forschungsantrag erneut dem Urteil der *Peers* unterwerfen. Aber erstens wird durch eine Schwerpunktverlagerung von *Peer Reviews* in den frühen Phasen einer Karriere die Abhängigkeit stark reduziert. Zweitens wird der Druck zur Produktion quantitativ messbaren Outputs gemildert und damit auch die Belastung der Gutachter. Drittens wird der Einfluss beherrschender Indizes abgebaut und mittelbar die Gefahr der Homogenisierung der Forschung durch das Regime der *Impact*-Faktoren gemindert.

(3) Die Finanzierung der Forschung ist wieder zu einem höheren Anteil durch Grundausstattungen und zu einem geringeren Teil durch Einwerbung von Drittmitteln zu gewährleisten: Auch diese Maßnahme verringert den Einfluss der *Peer Reviews*. Sie verhindert darüber hinaus das Entstehen ineffizienter und die Vielfalt reduzierender *research empires*. Zusätzlich wäre eine größere Dezentralisierung und Vielfalt von Fördereinrichtungen jenseits der dominierenden Deutschen Forschungsgemeinschaft wünschenswert.

Aus diesen Annahmen resultieren zwei Vorschläge, die über die Abschaffung von *pay for performance* und die verringerte Abhängigkeit vom *Peer Review*-System hinausgehen.

1. Das Ausmaß an regelmäßigen Evaluationen von Individuen ist zu reduzieren. Es schränkt die Autonomie ein, insbesondere wenn sie zu häufig und zu eng geführt werden. Der Aufwand ist groß, größer noch sind die verborgenen Kosten, die durch reaktives Verhalten der Evaluierten entstehen.

2. Die Evaluationen von Institutionen, die zur Verteilung von Ressourcen unvermeidlich sind, sind in erster Linie prozess- und nicht outputorientiert durchzuführen. Sie sollten sich dabei auf folgende Fragen konzentrieren: Ist ein sorgfältiger Prozess der Sozialisation und Selektion der Wissenschaftler gesichert? Ist ein hoher Grad an Autonomie im Forschungsprozess gewährleistet?

Ein nach diesen Gesichtspunkten gestaltetes Anreizsystem hat der deutschsprachigen Wissenschaft zu Beginn des 20. Jahrhunderts Weltgeltung verschafft. Ihr Ruf wird derzeit massiv durch ein Anreizsystem beeinträchtigt, das mit dem Charakter von Forschung unvereinbar ist. Die Anreize für kreative Forschung werden untergraben. Unsere Vorschläge können zwar nicht vollständig das verlorengegangene Vertrauen in die Selbststeuerungsfähigkeit der Wissenschaft wiederherstellen, aber sie können gute Voraussetzungen dafür schaffen, dass die Attraktivität des deutschen Wissenschaftssystems für eigenständige und originelle Forscher wieder zunimmt.