Rankings und der Preis der Wissenschaft[1]

Margit Osterloh und Bruno S. Frey

Wenn wir als Ökonom-innen vom Preis der Wissenschaft hören, denken wir zuerst an das Preissystem, das sich aus dem Zusammenspiel von Angebot und Nachfrage auf wettbewerblichen Märkten ergibt. Genau dieses liegt in der Wissenschaft nicht vor, auch wenn die Einführung des *New Public Managements* – das heißt, die Übertragung von Prinzipien der gewinnorientierten Privatwirtschaft – an Universitäten und Forschungsinstitutionen das Gegenteil suggerieren will. *New Public Management* möchte durch Output-Kontrolle, durch Leistungsmessung und daran geknüpfte leistungsabhängige Entlohnung ein Quasi-Preissystem einführen.

Universitäten haben die Prinzipien des *New Public Management* in den letzten Jahren mehr und mehr übernommen, weil sie glauben, die Öffentlichkeit mittels Zielvereinbarungen und Evaluationen über ihre Tätigkeit informieren zu müssen. Leistungen sollen anhand von Indikatoren für die Öffentlichkeit transparent und messbar gemacht werden. In den letzten Jahren hat der Drang, alles und jedes quantitativ zu erfassen, weite Teile der Gesellschaft erfasst. Auf der Strecke bleiben dabei qualitative Aspekte, die sich einer simplen Messung entziehen. Es ist bemerkenswert, dass sich auch eine Institution wie die Universität, die seit Jahrhunderten einen bedeutenden und unangefochtenen Platz in der Gesellschaft einnimmt, verunsichert fühlt und ihre Tätigkeit nach außen hin in Form quantitativer Faktoren rechtfertigen zu müssen glaubt.

Vor allem Rankings von Forschungsleistungen dienen dieser Quantifizierung. Rankings messen unter anderem die Anzahl von Publikationen in renommierten Zeitschriften sowie die Höhe der eingeworbenen Drittmittel. Sie sollen der Leistungsbeurteilung dienen und die Rolle von Preisen auf Märkten übernehmen. Anders als in der Wirtschaft gibt es aber in der Wissenschaft ein systematisches Marktversagen (vgl. Osterloh 2013). *Erstens* werden in der Wissenschaft öffentliche Güter produziert, die gekennzeichnet sind durch Nichtausschließbarkeit bei der Nutzung und Nichtrivalität im Konsum des produzierten Wissens. Forschung ist *zweitens* gekennzeichnet durch

1 Der vorliegende Text fasst von uns formulierte Argumente erstmals in deutscher Sprache zusammen und aktualisiert sie; vgl. Frey/Osterloh 2011 und 2014.

urn:nbn:de:hbz:6:3-2015051261

fundamentale Unsicherheit. Diese ist sichtbar an sogenannten Serendipitätseffekten, das heißt, Wissenschaftler_innen finden etwas anderes als das, was sie gesucht haben. Solche Effekte sind in der Wissenschaft zahlreich. Beispiele sind die Entdeckung des Dynamits, der Röntgenstrahlen oder der Radioaktivität. *Drittens* stellt sich der Nutzen wissenschaftlicher Entdeckungen mitunter erst nach sehr langer Zeit ein. In der Wissenschaft handelt es sich um sogenannte Vertrauensgüter im Unterschied zu Erfahrungsgütern. Bei letzteren kann man nach Gebrauch feststellen, ob sie etwas taugen oder nicht. Bei Vertrauensgütern ist dies nur langfristig oder manchmal nie möglich. Zum vierten gibt es Zurechnungsprobleme von Entdeckungen zu Personen. Die Wissenschaftsgeschichte ist voll von so genannten »Multiples« (Merton 1961). Damit werden Entdeckungen bezeichnet, die ursprünglich Einzelnen zugeschrieben wurden und die sich später als ›in der Luft liegend‹ herausgestellt haben, sodass nicht klar ist, wer der Entdecker war. Dazu gehört die Erfindung der Infinitesimalrechnung, die sowohl Leibniz als auch Newton für sich als Entdecker beanspruchen. Auch bei der Allgemeinen Relativitätstheorie existieren Zweifel, ob Einstein oder Hilbert der Erste war, oder in der Evolutionstheorie, ob es Darwin oder Wallace war (vgl. Simonton 2004).

Für den fehlenden Markt braucht die Wissenschaft einen Ersatz bei der Leistungsbeurteilung. Als solcher wird die Gelehrtenrepublik, die »Republic of Science« (Polanyi 1962/2002), angesehen. Innerhalb der Gelehrtenrepublik wird mittels Gutachten von wissenschaftlichen Kolleg_innen festgestellt, was gute Forschung ist und was nicht. Dazu dienen heute in erster Linie doppelt-blinde Begutachtungen, das heißt, Gutachten, bei denen die Gutachtenden und die Begutachteten einander nicht kennen oder kennen sollen. Dies soll durch eine Anonymisierung beider Seiten erreicht werden.

Taugt das doppelt-blinde Begutachtungs-Verfahren zur Leistungsbeurteilung?

Leider gibt es eine Fülle von empirischer Evidenz dafür, dass Begutachtungsverfahren im Allgemeinen, und doppelt-blinde Verfahren im Besonderen, mangelhaft funktionieren. *Erstens* gibt es eine geringe Übereinstimmung von Gutachterurteilen (vgl. Peters/Ceci 1982; Bornmann/Daniel 2003; Starbuck 2005). Die Korrelation zwischen Gutachterurteilen liegt zwischen 0.09 und 0.5.[2] Sie ist in den Naturwissenschaften keineswegs höher als in den Sozialwissenschaften (vgl. Nicolai/Schmal/Schuster 2015). Die Übereinstimmung von Gutachterurteilen ist im unteren Qualitätsbereich höher als im oberen Bereich (vgl. Moed 2007). In der klinischen Neurowissenschaft wurde sogar eine statistische Korrelation zwischen Gutachtenden festgestellt, die nicht signifikant höher war als die einer Zufallsauswahl (vgl. Rothwell/Martyn 2000). Die Auswahl der Gutach-

2 Die Korrelation von Gutachterurteilen für den Schweizer Nationalfonds wurde von Reinhart (2012) für die Fächer Biologie mit 0.45 und für die Medizin mit 0.2 in 1998 ermittelt. Die Korrelationen bei der Einschätzung der Qualität von Zeitschriften in der britischen ABS-Liste liegt mit durchschnittlich 0,68 höher, allerdings finden Peters et al. (2014) beträchtlichen *ingroup favoritism,* z.B. von US-amerikanischen Wissenschaftler_innen zugunsten US-amerikanischer Journals.

tenden hat einen entscheidenden Einfluss auf Annahme oder Ablehnung eines Papiers (vgl. Bornmann/Daniel 2009). *Zweitens* ist die prognostische Qualität von Gutachten gering. Die Reviewer-Einschätzungen korrelieren nur mit 0.25 bis 0.37 mit späteren Zitationen (vgl. Starbuck 2006). *Drittens* ist die zeitliche Konsistenz von Gutachterurteilen niedrig. Es gibt zahlreiche Beispiele dafür, dass in so genannten *A-journals* zurückgewiesene Artikel später berühmt wurden und Preise gewonnen haben, inklusive des Nobel-Preises (vgl. Gans/Shepherd 1994; Campanario 1996). Ein aktuelles Beispiel ist Daniel Shechtman, der Chemie-Nobelpreisträger des Jahres 2011. Er wurde gemäß Zeitungsberichten[3] für seine Entdeckung der Quasi-Kristalle zunächst von seinen Kollegen nicht nur ausgelacht, sondern auch aus seiner Forschungsgruppe hinaus geworfen. *Viertens* gibt es zahlreiche Bestätigungs-Fehler. Gutachtende fanden in 72 Prozent von Papieren methodische Fehler, wenn diese dem Mainstream widersprechen, hingegen nur in 25 Prozent der Fälle, wenn das Papier im »mainstream« (Mahoney 1977) liegend argumentiert. *Fünftens* gibt es einen beträchtlichen Institutionen- und Gender-Bias. Bei Forschungsanträgen favorisieren Gutachter Bewerbungen von prestigereichen Institutionen (vgl. Godlee/Gale/Martyn 1998). Der Nachweis eines Gender–Bias in Schweden bei der Vergabe von Forschungsgeldern hat viel Aufmerksamkeit erregt (vgl. Wenneras/Wold 1997). *Sechstens* erstellen anonyme Gutachtende oft sehr oberflächliche Berichte. Ihre Kommentare sind mitunter wenig hilfreich, vielmehr setzen sie die Autor_innen erheblich unter Druck. Das Ergebnis ist »Publication as Prostitution« (Frey 2003). *Siebtens* dauern Begutachtungsprozesse oft Monate, wenn nicht mehrere Jahre.

Und schließlich *achtens* ist das System unverhältnismäßig teuer: Steuerzahler_innen werden von den Zeitschriften-Verlagen gleich fünffach zur Kasse gebeten: Zum ersten zahlt der Staat Saläre für die Verfasser_innen der Artikel, zum zweiten für die Gutachtenden und Editor_innen, soweit diese ebenfalls an Universitäten beschäftigt sind. Zum dritten müssen heutzutage mitunter Beträge von 500-1500 US-Dollar aufgewendet werden, wenn man ein Papier einreicht. Zum vierten müssen die Universitätsbibliotheken Unsummen an Lizenzgebühren an eben diese Verlage entrichten, für die sie unentgeltlich schreiben, editieren und Gutachten erstellen. Schließlich müssen die Forschenden, wollen sie ihr veröffentlichtes Papier online stellen, häufig noch einmal eine Gebühr um die 1.000 US-Dollar oder oft auch mehr an die Verlage bezahlen.

Taugen Rankings als Unterstützung bei der Leistungsmessung?

Die Gelehrtenrepublik funktioniert nach diesen Befunden als Marktersatz nicht gut, obwohl sie immerhin Vorteile hat, nämlich Vieldimensionalität, Dezentralität und Vielfalt. Wird eine Publikation abgelehnt, kann man sie in anderen Journals ähnlicher Qualität einreichen. Auch herrscht im deutschsprachigen Universitätssystem eine große

3 Vgl. http://www.ftd.de/wissen/natur/:kopf-des-tages-daniel-shechtman-der-quasi-wissenschaftler/60112463.html (20.03.2015).

Vielfalt an Möglichkeiten zur Bewerbung an gleichwertigen Universitäten.[4] Dies bringt aber ein Problem mit sich: Die Öffentlichkeit, das heißt, Forschungsmanager_innen, Journalist_innen und Ministerien sind nicht in der Lage, mit einem einfachen Kriterium die Qualität der Forschung und der Forschenden zu beurteilen. Darauf aber habe die Öffentlichkeit – so die Botschaft des *New Public Managements* – einen Anspruch. Die Wissenschaft müsse über einfache und klare Kennzahlen rechenschaftspflichtig gegenüber dem Steuerzahler gemacht werden. Dazu dienen Publikations-Rankings. Rankings versprechen eine größere Objektivität als Peer Reviews, einen Ausgleich der Fehleinschätzungen der Gutachter_innen mittels Aggregation und ein Zurückdrängen der Altherren-Netzwerke. Können sie das leisten?

Rankings, darunter auch die meisten internationalen Universitätsrankings (vgl. Schmoch 2015), wie auch Individualrankings haben technische Probleme (vgl. Rost/Frey 2011; Osterloh 2012). Diese sind schwerwiegend, aber sie lassen sich mit hohen Kosten beheben. Wichtiger sind die unveränderbaren Probleme, welche durch paradoxe Effekte zustande kommen (vgl. Osterloh/Frey 2014). Diese können auf der individuellen und der institutionellen Ebene auftreten.

Auf der individuellen Ebene ergibt sich das größte Problem bei Rankings, die anhand der Veröffentlichungen der Forschenden in ›guten‹ oder so genannten *A-journals* erstellt werden. Dabei wird unterstellt, dass ein in einer ›guten Zeitschrift‹ veröffentlichter Artikel auch eine ›gute Publikation‹ darstellt, weil solche Zeitschriften die »kollektive Weisheit« (Laband 2013) einer *scientific community* darstellen. Was eine ›gute Zeitschrift‹ ist, wird meist durch den *impact factor* bestimmt, das heißt, einem Maß, wie oft im Durchschnitt alle Artikel in einer Zeitschrift im Zeitraum von zwei Jahren nach deren Veröffentlichung zitiert wurden. Diese Interpretation hat sich international durchgesetzt (vgl. Archambault/Larivière 2009, Jarwal, Brion/King 2009). Etwas anders gehen Rankings vor, welche Zeitschriften nach ihrer Reputation bewerten. Auch hier wird unterstellt, dass die Qualität eines einzelnen Aufsatzes nach der Reputation der Zeitschrift bemessen werden kann, in welcher der Aufsatz veröffentlicht wurde.

In beiden Fällen – Bewertung nach *impact factor* oder nach Reputation – ist dies aber ein unsinniges Kriterium. Wie inzwischen hinlänglich kritisiert (vgl. Oswald 2007; Baum 2010; Kieser 2012; Frey/Osterloh 2013), kann aus dem *impact factor* oder der Reputation einer Zeitschrift kein Rückschluss auf die Qualität eines *einzelnen* Artikels gezogen werden, der in dieser Zeitschrift veröffentlicht wurde. Einige wenige Aufsätze werden häufig zitiert; die allermeisten hingegen selten oder gar nicht.[5] Wer auch nur eine Grundausbildung in Statistik genossen hat, weiß, dass bei einer stark schiefen Verteilung Durchschnittswerte keine Aussagekraft haben. Gleichwohl verwenden Wissenschaftler_innen, die es eigentlich besser wissen müssten, diese Art der Qualitätsbewertung bei der Entscheidung über die Karrieren von Nachwuchskräften. Vielfach ist eine

4 Kritischer sind allerdings die Auswirkungen bei der Begutachtung von Forschungsanträgen, weil in Deutschland eine starke Konzentration der zu vergebenden prestigereichen Drittmittel bei einer Institution besteht, der DFG.

5 Dies ist ein weiterer Beleg für die Unzuverlässigkeit von Peer Reviews. Auch Gutachtende für *top-journals* sind offensichtlich nicht in der Lage, die Relevanz eines Artikels richtig einzuschätzen, damit ein ausgeglichenes Bild der Zitationen entsteht.

Habilitation weitgehend Formsache, wenn entsprechend diesen Kriterien genügend Publikationen in so genannten *A-journals* erreicht werden. Ganz ähnlich wird bei Berufungen auf Professuren vorgegangen. Einige Universitäten zahlen auch noch Geldbeträge für Publikationen in ›guten‹ *journals*.[6]

Die Einsicht, dass die Veröffentlichung in einem ›guten‹ *journal* nicht gleichzusetzen ist mit einer ›guten‹ Publikation, setzt sich langsam, aber stetig durch. Gemäß der *International Mathematical Union IMU* (2008) ist die Wahrscheinlichkeit, dass ein zufällig ausgewählter Artikel in einer Zeitschrift mit einem niedrigen *impact factor* zitiert wird, um 62 Prozent höher ist als in einer Zeitschrift mit einem fast doppelt so hohem *impact factor.* Man irrt somit in 62% der Fälle, wenn man sich nach dem *impact factor* richtet! Der Schweizerische Nationalfonds hat jüngst die *DORA-Deklaration (San Francisco Declaration of Research Assessment)* unterschrieben. Danach darf die Qualität eines Aufsatzes nicht nach dem *impact factor* der veröffentlichenden Zeitschrift bewertet werden (vgl. DORA, 2012). Der Chefredaktor von *Science,* Bruce Alberts, stellt in einem im Mai 2013 publizierten Leitartikel unmissverständlich fest: »As frequently pointed out by leading scientists, this impact factor mania makes no sense [...]. Such metrics [...] block innovation« (Alberts 2013: 787). Der Grund dafür ist nicht nur die hohe Fehlerwahrscheinlichkeit bei der Beurteilung von Artikeln gemäß *impact factor* oder Reputation der Zeitschrift, sondern auch die Gefahr, dass die intrinsische Motivation der Forschenden reduziert wird (vgl. Frey/Osterloh 2002; Osterloh/Frey 2014).

Diese Probleme werden auf institutioneller Ebene durch *lock-in*-Effekte verstärkt, welche zu einer sich selbst erfüllenden Prophezeiung führen. Bei diesen Voraussagen werden die Bedingungen geändert, unter denen die Voraussage erfüllt wird (vgl. Merton 1948). Diesen Effekten können sich einzelne Personen oder Institutionen nur schwer entziehen, auch wenn sie deren Schädlichkeit erkennen. Beispiele hierfür sind:

- Wissenschaftler_innen beugen sich dem Druck des *publish or perish*-Diktats. Das Publikations-Rad dreht sich immer schneller, ohne dass der Wettbewerb zu einer höheren Qualität von Papieren und zur besseren Dissemination von darin publizierten Erkenntnissen führt (vgl. Laband/Tollison 2003). Nur 50 Prozent der in referierten Zeitschriften veröffentlichten Artikel werden von anderen als den Autoren und den Gutachtern gelesen, und 90 Prozent der Artikel werden niemals zitiert (vgl. Meho 2006). Darüber hinaus scheint die Zuverlässigkeit der empirischen Forschung zu sinken (vgl. Gläser et al. 2008).
- An britischen Universitäts-Departments werden vor dem Termin des *Research Assessment Exercise* Forscher_innen mit hohen Publikationszahlen teuer eingekauft. Diese veredeln zwar die Publikations- und Zitationsindikatoren, sind aber in der Regel kaum an den jeweiligen Universitäten anwesend.
- Zeitschriften-Herausgeber_innen fordern ihre Autor_innen auf, mindestens vier bis fünf Artikel ihrer Zeitschrift zu zitieren, um ihren *impact-factor* zu erhöhen.

6 Selbstverständlich haben Artikel in einem *A-journal* eine besonders hohe Chance, zur Kenntnis genommen und zitiert zu werden. Deshalb müssten eigentlich die Zitationen von Autor_innen in einem *B-* und *C-journal* höher und die von Autor_innen in einem *A-journal* niedriger bewertet werden (vgl. Balaban 2012).

– Fakultäten berufen neue Mitglieder in erster Linie gemäß ihrem Publikationsranking, um ihre ›Leuchtturm-Position‹ zu stärken.
– Befunde aus Großbritannien und Australien mit ihren stark entwickelten Systemen der evaluationsbasierten Ressourcenzuweisung zeigen, dass dort die Forschung homogener geworden ist und sich mehr am *mainstream* orientiert (vgl. Lee 2007) und dass die Etablierung von neuen Forschungsfeldern schwieriger geworden ist (vgl. Hargreaves Heap 2002).

Die Ursachen solcher *lock-in*-Effekte werden von Espeland/Sauder (2007) analysiert. Zum ersten vergrößern Rankings die Unterschiede unverhältnismäßig. Die Unterschiede zwischen den Ranking-Positionen mögen noch so gering sein, in der Wahrnehmung der Öffentlichkeit bekommen sie ein großes Gewicht. Entsprechend hart ist der Kampf um die Positionen. Auf diese Weise werden die vielfältigen Kriterien, nach denen eine komplexe Leistung beurteilt werden müsste, durch Rankings zunehmend in eine einfältige, hierarchische Rangordnung gebracht. Diese untergräbt den »institutionalisierten Skeptizismus« (Merton 1973), der gute Forschung ausmacht.

Dies bewirkt zum zweiten, dass Ressourcen innerhalb der Universitäten anders verteilt werden. Für Rankings relevante Aktivitäten werden verstärkt zulasten von Aktivitäten mit geringer Sichtbarkeit. Beispiele sind Investitionen ins Universitäts-Marketing, um sich als ›Leuchtturm‹ zu etablieren.

Drittens entstehen »Matthäus-Effekte« im Sinne des »Wer hat, dem wird gegeben« (Merton 1968). Mittel werden bevorzugt an diejenigen verteilt, die in der Vergangenheit die Rankings anführten, wodurch deren Reputation und auch Leistungsfähigkeit entsprechend der Ranking-Kriterien gesteigert wird. Dies erklärt, warum vergangene Rankings die stärksten Prädiktoren für gegenwärtige Rankings sind (vgl. Stake 2006). Dies geschieht ungeachtet der Tatsache, dass es auch in der Forschung einen abnehmenden Grenznutzen von Ressourcen gibt (zu empirischen Befunden vgl. Jansen et al. 2007), weshalb mehr Ressourcen für exzellente Forscher_innen oder Forschungsgruppen nicht immer effizient sind. Vielmehr kann es einen höheren Zusatznutzen erbringen, zunächst mittelmäßige Forschung zu fördern.

Viertens werden Neigungen gefördert, die Regeln zu manipulieren, um das ›System zu schlagen‹. Dies ist um so stärker zu erwarten, umso mehr die intrinsische Motivation eines »taste for science« (Merton 1973) zerstört ist. Beispiele hierfür sind Hochschulen, die schlechte Studierende in Sondergruppen abschieben (z.B. in Klassen für vorläufig Aufgenommene), welche nicht in der Statistik erscheinen (vgl. Gioia/Corley 2002).

Dies verweist auf das »Performance Paradox« (Meyer/Gupta 1994; Meyer 2009; Frost/Brockmann 2014). Alle Leistungsindikatoren haben die Tendenz, mit der Zeit ihre Relevanz zu verlieren. In der Folge können nicht mehr gute von schlechten Leistungen anhand der Indikatoren unterschieden werden. Die Ursache sind zwei gegenläufige Effekte, die allerdings in der Realität nur schlecht auseinander gehalten werden können: Leistungsindikatoren können einerseits einen positiven Lerneffekt hervorrufen, zum Beispiel deutlich machen, dass für gute Wissenschaft Publikationen ausschlaggebend sind. Dies kann positive Anreiz-, Selektions- und Selbst-Selektions-Effekte bewirken, wodurch die Varianz der Leistung sinkt. Andererseits kann dieses Sinken auch eine ganz andere Ursache haben, nämlich einen perversen Lerneffekt. Dieser tritt dann auf, wenn

der Fokus auf die Leistungsindikatoren gelegt wird und nicht auf das, was er messen soll: »When a measure becomes a target, it ceases to be a good measure« (Strathern 1996: 4).

Die einzige Methode, um diesem Paradox zu entrinnen und dennoch Leistungsindikatoren beizubehalten, wäre deren ständige Veränderung und Anpassung durch die betroffenen Fachleute. Dagegen werden diejenigen Protest anmelden, welche mit Hilfe der bestehenden Indikatoren Einfluss errungen haben. Dazu gehört auch das Wissenschafts-Management, das in Universitäten und Forschungseinrichtungen durch eine »Governance by Numbers« (Heintz 2008) gegenüber den Forschenden an Einfluss gewonnen hat. Es beansprucht immer höhere Anteile am Forschungsbudget,[7] obwohl dies der wissenschaftlichen Leistung der Institution wenig nützt (vgl. Goodall 2009; Goodall/Bäker 2015), sondern stattdessen die Forschungsbürokratie stärkt.

Alternativen?

Wie kann man angesichts der geschilderten Situation dennoch eine Leistungsbewertung vornehmen? Der erste Vorschlag besteht in der Reduktion der Anlässe für Evaluationen auf wenige karrierererelevante Entscheidungen, zum Beispiel bei der Bewerbung um eine Stelle oder bei der Beantragung von zusätzlichen Forschungsmitteln. Eine sorgfältige Eingangskontrolle ersetzt die kontinuierliche Bewertung durch dauernde Evaluationen (vgl. Osterloh 2010; Frey/Osterloh 2012; Frey/Homberg/Osterloh 2013; Osterloh/Frey 2014). Sie hat die Aufgabe, das Innovationspotential, die Motivation für selbstorganisiertes Arbeiten und die Identifikation mit dem „taste of science“ (Merton 1973) zu überprüfen. Wer dieses ›Eintrittsticket‹ in die Gelehrtenrepublik aufgrund einer rigorosen Prüfung erworben hat, sollte weitgehende Autonomie einschließlich einer angemessenen Grundausstattung erhalten. Eine solche Eingangskontrolle ist keineswegs neu. Sie wird an den *Institutes for Advanced Studies* ebenso praktiziert wie an der Harvard-Universität, in deren Prinzipien es heißt: »The primary means for controlling the quality of scholarly activities of this faculty is through the rigorous academic standards applied in selecting its members«.[8] Dieses Konzept hilft, die geschilderten Schwächen der Begutachtungsprozesse zu reduzieren, weil Begutachtungen auf wenige Anlässe beschränkt werden. Die unbeabsichtigten Nebenwirkungen und *Ranking Games* in der Forschung werden reduziert. Das Konzept ist aber gleichwohl auf Gutachten mit allen geschilderten Problemen angewiesen, auch wenn diese weniger häufig erforderlich sind.

Hier verspricht der zweite Vorschlag Abhilfe: das offene *post-publication-peer-review*-Verfahren (vgl. Kriegeskorte 2012; Frey/Osterloh 2014; Osterloh/Kieser 2015). Dieses Verfahren sieht widersprüchliche Gutachten nicht als Problem, sondern als ein

7 Der Europäische Rechnungshof warf der EU im Jahr 1997 im Zusammenhang mit dem 4. Forschungsrahmenprogramm »aufgeblähte Bürokratie und sinnlose Geldverschwendung« vor. Von den zur Verfügung stehenden 13 Milliarden kamen nur etwa 60 Prozent bei den Forschungsinstitutionen an (vgl. Binswanger 2010: 178).

8 http://www.fas.harvard.edu/research/greybook/principles.html.(20.03.2015).

Zeichen solider und produktiver Wissenschaft. Kontroversen bieten Anlass für die Fortentwicklung der Wissenschaft. Dies ist allerdings nur dann der Fall, wenn Gutachten zu einem offenen wissenschaftlichen Diskurs führen, was bei der derzeitigen anonymen Doppelt-Blind-Begutachtung nicht möglich ist.

Im neuen Verfahren würden Forschende eine_n erfahrene_n Kolleg_in als *Editor* beauftragen, Kommentare einzuholen, welche auf einer gemeinsamen Plattform veröffentlicht werden. Die Stellungnahmen sollten mit Namen gekennzeichnet sein und können als kleine zitierfähige und reputationswirksame Veröffentlichungen gelten. Die Verfasser_innen des ursprünglichen Artikels können auf derselben Plattform antworten. Nur wenn ein lebendiger Diskurs zustande kommt, ist der ursprüngliche Aufsatz wissenschaftlich ergiebig. Erhält ein Papier keinen oder wenige Kommentare, signalisiert dies mangelhafte Qualität bzw. wissenschaftliche Irrelevanz. Sind die Kommentare oberflächlich oder gar feindselig (wie dies bei anonymen Gutachten allzu häufig der Fall ist), schädigt dies die Reputation des Gutachtenden. Die Transparenz schafft einen Anreiz, fundierte Einschätzungen zu schreiben. Nach einiger Zeit könnten diejenigen Beiträge, welche die lebhaftesten Diskussionen ausgelöst haben, als *state of the art* in elektronischen Sammelwerken ausgewiesen werden.

Dieses neue System würde das Begutachtungssystem endlich in das Internetzeitalter führen. Es beseitigt das Platzproblem, weil im Internet unbeschränkt viel Raum für Publikationen zur Verfügung steht.[9] Es kann viel schneller arbeiten als das träge heutige Begutachtungssystem, bei dem mitunter zwei bis drei Jahre von der Einreichung bis zur Veröffentlichung verstreichen. Bei interessanten Papieren wäre eine rasche Rückkopplung zu erwarten. Darüber hinaus erspart es Steuerzahler_innen die immensen Kosten, welche ihnen die Verlage heute auferlegen. Das Verfahren verführt vor allem deutlich weniger zu einem *gaming the system*. Entscheidend ist jedoch: Argumentativer Diskurs in der *Republic of Science* erhält wieder Vorrang gegenüber Zählübungen wie *impact factors* und Rankings.

Die Durchsetzung dieses neuen Verfahrens wäre nicht einfach. Neben Gewinner_innen (dem wissenschaftlichen Nachwuchs) gibt es auch Verlierer_innen (vor allem Verlage). Auch hier dürften Einrast- oder *lock-in*-Effekte eintreten, die den Übergang erschweren. Aber angesichts der riesigen Probleme des heutigen Systems wäre zu wünschen, dass neue Alternativen aufgezeigt und ernsthaft diskutiert werden.

Unser Beitrag kritisiert die heute in der Wissenschaft Rankings zugewiesene Rolle aus der Perspektive von Ökonomen. Aus dieser Perspektive sind Rankings nicht in der Lage, ein System von Märkten und Preisen künstlich herzustellen und so den Preis der Wissenschaft zu bestimmen. In der Wissenschaft herrscht (sinnvollerweise) Marktversagen. Wir zeigen gleichzeitig auf, dass es Alternativen der Leistungsbeurteilung gibt, welche das intrinsische Interesse an der Forschung fördern, und die gleichzeitig Freiräume schaffen, die eine lebendige und kreative Wissenschaft braucht.

9 Allerdings darf angenommen werden, dass der Publikationsdruck in dem neuen Verfahren nachlässt, weil die Qualität des ausgelösten Diskurses und nicht die Anzahl der Veröffentlichungen zählt.

Literatur

Alberts, Bruce (2013): »Editorial: Impact Factor Distortions«. In: *Science* 340, 787.

Archambault, Éric/Larivière, Vincent (2009): »History of the Journal Impact Factor: Contingencies and Consequences«. In: *Scientometrics* 79: 3, 639-653.

Balanban, Alexandru T. (2012): »Positive and Negative Aspects of Citation Indices and Journal Impact Factors«. In: *Scientometrics* 92, 241-247.

Baum, Joel A. C. (2010): »Free-Riding on Power Laws: Questioning the Validity of the Impact Factor as a Measure of Research Quality in Organization Studies«. In: *Organization* 18: 4, 449-466.

Binswanger, Mathias (2010): *Sinnlose Wettbewerbe. Warum wir immer mehr Unsinn produzieren,* Freiburg/Breisgau: Herder.

Bornmann, Lutz/Daniel, Hans-Dieter (2003): »Begutachtung durch Fachkollegen in der Wissenschaft. Stand der Forschung zur Reliabliliät, Fairness und Validität des Peer-Review-Verfahrens«. In: *Universität auf dem Prüfstand. Konzepte und Befunde der Hochschulforschung,* hg. v. Stefanie Schwarz/Ulrich Teichler, Frankfurt/Main: Campus, 211-230.

Bornmann, Lutz/Daniel, Hans-Dieter (2009): »The Luck of the Referee Draw: The Effect of Exchanging Reviews«. In: *Learned publishing* 22: 2, 117-125.

Campanario, Juan Miguel (1996): »Using Citation Classics to Study the Incidence of Serendipity in Scientific Discovery«. In: *Scientometrics* 37, 3-24.

Dora (San Francisco Declaration on Research Assessment) (2012): *Accessed December* 16, 2012. http://am.ascb.org/dora/files/SFDeclarationFINAL.pdf.

Espeland, Wendy Nelson/Sauder, Michael (2007): »Rankings and Reactivity: How Public Measures Recreate Social Worlds«. In: *American Journal of Sociology* 113: 1, 1-40.

Frey, Bruno S. (2003): »Publishing as Prostitution? – Choosing between One's Own Ideas and Academic Success«. In: *Public Choice* 116, 205-223.

Frey, Bruno S./Homberg, Fabian/Osterloh, Margit (2013): »Organizational Control Systems and Pay-for-Performance in the Public Service«. In: *Organization Studies* 34: 7, 949-972.

Frey, Bruno S./Osterloh, Margit (2002): *Managing Motivation,* 2. Aufl., Wiesbaden: Gabler.

Frey, Bruno S./Osterloh, Margit (2011): »Rankings Games«. *University of Zurich. Department of Economics, Working Paper* No. 39, http://ssrn.com/abstract=1957162 (17.02.2015).

Frey, Bruno S./Osterloh, Margit (2012): »Rankings: Unbeabsichtigte Nebenwirkungen und Alternativen«. In: *Ökonomenstimme,* 17.02.2012, http://www.oekonomenstimme.org/artikel/2012/02/rankings-unbeabsichtigte-nebenwirkungen-und-alternativen/ (17.02.2015).

Frey, Bruno S./Osterloh, Margit (2013): »Gut publizieren = gute Publikation?«. In: *Ökonomenstimme,* 16.5.2013, http://www.oekonomenstimme.org/artikel/2013/05/gut-publizieren--gute-publikation/ (17.02.2015).

Frey, Bruno S./Osterloh, Margit (2014): »Schlechte Behandlung des wissenschaftlichen Nachwuchses und wie man das ändern könnte«. In: *Ökonomenstimme,* 28.10.2014, http://www.oekonomenstimme.org/artikel/2014/10/schlechte-behandlung-des-wissenschaftlichen-nachwuchses-und-wie-man-das-aendern-koennte/ (17.02.2015).

Frost, Jetta/Brockmann, Julia (2014): »When Quality is Equated with Quantitavite Productivity – Scholars Caught in a Performance Paradox«. In: *Zeitschrift für Erziehungswissenschaft* 17: 6, Supplement, 25-45.

Gans, Joshua S./Shepherd, George B. (1994): »How are the Mighty Fallen: Rejected Classic Articles by Leading Economists«. In: *Journal of Economic Perspectives* 8, 165-179.

Gioia, Dennis A.,/Corley, Kevin. G. (2002): »Being Good versus Looking Good: Business School Rankings and the Circean Transformation from Substance to Image«. In: A*cademy of Management Learning and Education* 1, 107-120.

Gläser, Jochen et al. (2008): »Evaluationsbasierte Forschungsfinanzierung und ihre Folgen«. In: *Wissensproduktion und Wissenstransfer. Wissen im Spannungsfeld von Wissenschaft, Politik und Öffentlichkeit,* hg.v. Renate Mayntz et al., Bielefeld: transcript, 145-170.

Godlee, Fiona/Gale, Catharine. R./Martyn, C. N. (1998): »The Effect on the Quality of Peer Review of Blinding Reviewers and Asking them to Sign their Reports. A Randomised Controlled Trial«. In: *Journal of the American Medical Association* 263: 10, 1438-1441.

Goodall, Amanda H. (2009): »Highly Cited Leaders and the Performance of Research Universities«. In: *Research Policy* 38, 1070-1092.

Goodall, Aamanda. H./Bäker, Agnes (2015): »A Theory Exploring How Expert Leaders Influence Performance in Knowledge-Intensive Organizations«. In: *Incentives and Performance - Governance of Research Organization,* hg. v. Isabell M. Welpe et al., Heidelberg: Springer, 49-68.

Hargreaves Heap, Shaun P. (2002): »Making British Universities Accountable«. In: *Science Bought and Sold: Essays in the Economics of Science,* hg. v. Philip Mirowski/Esther-Mirjam Sent, Chicago: University of Chicago Press, 387-411.

Heintz, Bettina (2008): »Governance by Numbers. Zum Zusammenhang von Quantifizierung und Globalisierung am Beispiel der Hochschulpolitik«. In: *Governance von und durch Wissen,* hg. v. Gunnar Folke Schuppert/Andreas Voßkuhle, Baden-Baden: Nomos, 110-128.

International Mathematical Union Imu (2008): *Citation Statistics. A Report. Corrected version,* 16/12/08, http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf.

Jansen, Dorothea et al. (2007): »Drittmittel als Performanzindikator der wissenschaftlichen Forschung. Zum Einfluss von Rahmenbedingungen auf Forschungsleistung«. In: *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 59, 125-149.

Jarwal, S. D./Brion, A. M./King, M. L. (2009): »Measuring Research Quality Using the Journal Impact Factor, Citations and ›Ranked Journals‹: Blunt Instruments or Inspired Metrics?« In: *Journal of Higher Education Policy and Management* 31: 4, 289-300.

Kieser, Alfred (2012): »Jourqual – der Gebrauch, nicht der Missbrauch, ist das Problem. Oder: Warum Wirtschaftsinformatik die beste deutschsprachige betriebswirtschaftliche Zeitschrift ist«. In: *Die Betriebswirtschaft* 72, 93-110.

Kriegeskorte, Nikolaus (2012): »Open Evaluation: a Vision for Entirely Transparent Post-Publication Peer Review and Rating for Science«. In: *Frontiers in Computational Neuroscience* 6, 1-18.

Laband, David N. (2013): »On the Use and Abuse of Economics Journal Rankings«. In: *The Economic Journal* 123: 570, F223-F254.

Laband, David N./Tollison, Robert C. (2003): »Dry Holes in Economic Research«. In: *Kyklos* 56, 161-174.

Lee, Frederic S. (2007): »The Research Assessment Exercise, the State and the Dominance of Mainstream Economics in British Universities«. In: *Cambridge Journal of Economics* 31, 309-325.

Mahoney, Michael J. (1977): »Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System«. In: *Cognitive Therapy Research* 1, 161-175.

Meho, Lokman. I. (2006): *The Rise and Fall of Citation Analysis. Preprint physics.* http://arxiv.org/pdf/physics/0701012 (17.02.2015).

Merton, Robert K. (1948): »The Self-Fulfilling Prophecy«. In: *Antioch Review* 8, 193-210.

Merton, Robert K. (1961): »Singletons and Multiples in Scientific Discovery: A Chapter in the Sociology of Science«. In: *Proceedings of the American Philosophical Society* 105: 5, 470-486.

Merton, Robert K. (1968): »The Matthew Effect in Science«. In: *Science* 159, 56-63.

Merton, Robert K. (1973): *The Sociology of Science: Theoretical and Empirical Investigation,* Chicago: University of Chicago Press.

Meyer, Marshall W. (2009): *Rethinking Performance Management. Beyond the Balanced Scorecard,* Cambridge: Cambridge University Press.

Meyer, Marshall W./Gupta, Vipin. (1994): »The Performance Paradox«. In: *Research in Organizational Behavior* 16, 309-369.

Moed, Henk F. (2007): »The Future of Research Evaluation Rests with an Intelligent Combination of Advanced Metrics and Transparent Peer Review«. In: *Science and Public Policy* 34, 575-583.

Nicolai, Alexander T./Schmal, Stanislaw/Schuster, Charlotte L. (2015): »Interrater Reliability of the Peer Review Process in Management Journals«. In: *Incentives and Performance – Governance of Research Organizations,* hg. v. Isabell M. Welpe et al., Heidelberg: Springer, 107-120.

Osterloh, Margit (2010): »Governance by Numbers. Does it Really Work in Research?«. In: *Analyse/Kritik* 32, 267-283.

Osterloh, Margit (2012): »›New Public Management‹ versus ›Gelehrtenrepublik‹. Rankings als Instrument der Qualitätsbeurteilung in der Wissenschaft?« In: *Hochschule als Organisation,* hg. v. Uwe Wilkesmann/Christian J. Schmid, Wiesbaden: VS Verlag für Sozialwissenschaften/Springer Fachmedien, 209-221.

Osterloh, Margit (2013): »Das Paradox der Leistungsmessung und die Nachhaltigkeit der Forschung«. In: *Nachhaltigkeit in der Wissenschaft,* hg. v. Jörg Hacker, Nova Acta Leopoldina NF 117: 398, 103-113.

Osterloh, Margit/Frey, Bruno S. (2014): *Ranking Games. Evaluation Review,* online DOI: 10.1177/0193841X14524957.

Osterloh, Margit/Kieser, Alfred (2015): »Double-Blind Peer Review: How to Slaughter a Sacred Cow«. In: *Incentives and Performance – Governance of Research Organizations,* hg. v. Isabell M. Welpe et al., Heidelberg: Springer, 307-324.

Oswald, Andrew J. (2007): »An Examination of the Reliability of Prestigious Scholarly Journals: Evidence and Implications for Decision Makers«. In: *Economica* 74, 21-31.

Peters, Douglas P./Ceci, Stephen J. (1982): »Peer Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again«. In: *The Behavioral and Brain Sciences* 5, 187-195.

Polanyi, Michael (1962): »The Republic of Science: Its Political and Economic Theory«. In: *Minerva* 1, 54-73. Wieder abgedruckt in: Mirowski, Philip/Sent, Esther-Mirjam (2002): *Science Bought and Sold. Essays in the Economics of Science,* Chicago: The University of Chicago Press, 465-485.

Reinhart, Martin (2012): *Soziologie und Epistemologie des Peer Reviews,* Baden-Baden: Nomos.

Rost, Katja/Frey, Bruno S. (2011): »Quantitative and Qualitative Rankings of Scholars«. In: *Schmalenbach Business Review* 63, 63-91.

Rothwell, P. M./Martyn, C. N. (2000): »Reproducibility of Peer Review in Clinical Neuroscience. Is Agreement between Reviewers Any Greater than would be Expected by Chance Alone?«. In: *Brain* 123, 1964-1969.

Schmoch, Ulrich. (2015): »The Informative Value of International University Rankings: Some Methodological Remarks«. In: *Incentives and Performance – Governance of Research Organizations,* hg. v. Isabell M. Welpe et al., Heidelberg: Springer, 141-154.

Simonton, Dean Keith (2004): *Creativity in Science. Chance, Logic, Genius, and Zeitgeist,* Cambridge: Cambridge University Press.

Stake, Jeffrey Evans (2006): »The Interplay between Law School Rankings, Reputations, and Resource Allocations: Ways Rankings Mislead«. In: *Indian Law Journal* 82, 229-270.

Starbuck, William H. (2005): »How Much Better are the Most Prestigious Journals? The Statistics of Academic Publication«. In: *Organization Science* 16, 180-200.

Starbuck, William H. (2006): *The Production of Knowledge. The Challenge of Social Science Research,* Cambridge: Oxford University Press.

Strathern, Marylin (1996): »From Improvement to Enhancement: An Anthropological Comment on the Audit Culture«. In: *Cambridge Anthropology* 19, 1-21.

Wenneras, Christine/Wold, Agnes (1999): »Bias in Peer Review of Research Proposals in Peer Reviews in Health Sciences«. In: *Peer Review in Health Sciences,* hg. v. Fiona Godlee/Tom Jefferson, London: BMJ Books, 79-89.