CREMA

Center for Research in Economics, Management and the Arts

# Würfelt Gott? Würfelt die Wissenschaft?

Working Paper No. 2015-19

CREMA Südstrasse 11 CH - 8008 Zürich www.crema-research.ch

Margit Osterloh [1] 10.12.2015

# Würfelt Gott? Würfelt die Wissenschaft?

„Gott würfelt nicht“. Diese Überzeugung von Albert Einstein hat sich in der Quantenmechanik nicht durchgesetzt (Hattrupp, 2011). Heute geht man davon aus, dass auch bei physikalischen Zuständen Zufall eine Rolle spielt. Inwieweit gehört Zufall generell zur Wissenschaft bzw. zur Beurteilung der wissenschaftlichen Qualität? Wie viel Zufall gibt es hier bzw. sollte es geben?

Das wissenschaftliche Qualitätsbeurteilungs-System ist in der letzten Zeit unter Beschuss gekommen. Zunehmend wird angezweifelt, dass Impact-Faktoren, Journal-Rankings oder Anzahl von Zitationen ein angemessenes Kriterium für die Qualität wissenschaftlicher Forschung darstellen (z.B. Frey & Osterloh, 2014; 2016). Ebenso wird das wissenschaftliche Begutachtungsverfahren durch Peers - als der Königsweg der Qualitätsbeurteilung innerhalb der „Gelehrtenrepublik“ mittlerweile in Frage gestellt (z.B. Bornmann, 2011; Osterloh, 2015). „Peer review is.....largely a lottery, anti-innovatory, slow, expensive, wasteful of scientific time, inefficient, easily abused, prone to bias, unable to detect fraud and irrelevant“ (Smith, 2015). So lautet das Urteil des ehemaligen Editors der einflussreichen „British Medical Journals“. Die Annahme oder Ablehnung eines Artikels durch ein Journal hängt deutlich vom Glück ab, dem „richtigen“ Gutachter zugeteilt zu werden (Bornmann & Daniel, 2009). Es gibt zahlreiche Beispiele dafür, dass in so genannten A-Journals zurückgewiesene Artikel später berühmt wurden (Siler, Lee & Bero, 2015) oder dass Erfindungen auf sogenannte Serendipitäts-Effekte zurückgehen (Simonton, 2004), d.h. eine starke Zufallskomponente enthalten. Zufall ist demnach – wie in der Quantenphysik - in der wissenschaftlichen Qualitätsbeurteilung von Bedeutung.

Zufallsverfahren gelten jedoch als nicht rational, was wohl auch Einstein zu seiner Überzeugung geführt haben mag, dass Gott nicht würfelt (Davies, 1995). Allerdings haben solche Verfahren in der Geschichte eine herausragende Rolle gespielt, z.B. bei der Auswahl von Regierenden im alten Athen oder von Professoren an der Universität Basel im 18. Jahrhundert (Stolz, 1986). Ich möchte zeigen, dass in Zufallsverfahren mehr

[1] Prof. Dr. Dr.h.c. Margit Osterloh ist Professorin (em.) an der Universität Zürich, Forschungsdirektorin am Center for Research in Economics, Management and the Arts (CREMA) und ständige Gastprofessorin an der Universität Basel am Center for Research in Economics and Well-Being (CREW).

Rationalität steckt als gemeinhin angenommen, weshalb man durchaus aus der Not von zufallsabhängigen Ergebnissen in der Wissenschaft – methodisch angewendet – eine Tugend machen kann.

Welche Vorteile hat das fast vergessenes Entscheidungsverfahren des Zufalls und wie können diese in der Qualitätsbeurteilung in der Forschung eingesetzt werden? Warum sollten wir uns überhaupt Gedanken um ein so merkwürdig erscheinendes Verfahren machen? Die Antwort liegt darin, dass Qualitätsbeurteilung in der Wissenschaft außerordentlich schwierig ist.

# 1 Warum ist Qualitätsbeurteilung in der Wissenschaft schwierig?

In der Wissenschaft versagt der Marktmechanismus, dem wir in unserer Wirtschaftsordnung ansonsten eine hohe Verlässlichkeit und Objektivität zuschreiben. Deswegen müssen hier andere Entscheidungs-Verfahren zur Anwendung kommen, insbesondere das Gutachter- oder „peer review"-Verfahren. [2]. Was sind die Gründen des Marktversagens in der Wissenschaft? Es sind dies Marktversagen, Probleme der Peer Reviews sowie Probleme von Kennzahlen und Rankings.

## 1.1 Marktversagen

Trotz aller Bemühungen des New Public Managements, Leistungen durch den Markt bewerten zu lassen, gelingt dies in der Wissenschaft nicht (Osterloh & Frey, 2015; Frey & Osterloh, 2016). Verursacht ist das durch vier besondere Merkmale wissenschaftlicher Arbeit.

Zum *ersten* produziert Wissenschaft öffentliche Güter, gekennzeichnet durch Nichtausschließbarkeit bei der Nutzung sowie Nichtrivalität im Konsum der produzierten Wissens.

Zum *zweiten* ist Forschung gekennzeichnet durch fundamentale Unsicherheit. Diese ist sichtbar an sogenannten Serendipitäts-Effekten, d.h. daß man etwas anderes findet als das, wonach man gesucht hat. Solche Effekte sind in der Wissenschaft

[2] Zu einer Übersicht über verschiedene idealtypische Entscheidungsverfahren (Märkte/Preise, Hierarchie, Demokratie, Verhandlungen/Kompromiss, Konsens, Tradition, Zufall) und ihre Anwendungmöglichkeiten siehe Frey & Pommerehne (1990); Frey & Kirchgässner (1994); Frey & Steiner (2014) Das peer review-Verfahren wäre im Idealfall dem Konsens zuzurechnen.

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 werden sogenannte Vertrauensgüter im Unterschied zu Erfahrungsgütern produziert. Bei letzteren kann man nach Gebrauch feststellen, ob sie etwas taugen oder nicht. Bei Vertrauensgütern ist dies nur sehr langfristig oder manchmal nie möglich. Beispielsweise dauert die Umsetzung von Grundlagenforschung in praktische Anwendungen in der Medizin im Durchschnitt drei Jahrzehnte (Contopoulos-Ioannidis et al., 2008).

V*iertens* gibt es Zurechnungsprobleme von Entdeckungen zu Personen. Die Wissenschaftsgeschichte ist voll von sogenannten Multiples (Merton, 1961), d.h. Entdeckungen, 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, bei der nicht klar ist, ob Leibnitz oder Newton sie zuerst erfunden haben. Ebenso ist es mit der Allgemeinen Relativitätstheorie, bei der Zweifel existieren, ob Einstein oder Hilbert der Entdecker war, oder mit der Evolutionstheorie mit Darwin versus Wallace (z.B. Simonton, 2004).

Für den fehlenden Markt braucht Wissenschaft einen Ersatz. Das ist die Gelehrtenrepublik, die „Republic of Science" oder die „Scientific Community". Diese stellt mit Gutachten („peer reviews") fest, was gute Forschung ist. Das bringt das Zitat von Polanyi (1962/2002: 479) zum Ausdruck: „The soil of academic science must be exterritorial in order to secure its rule by scientific opinion" . Die Qualität der Forschung ist also nur von „innen heraus" durch die „Scientific Community" mittels „peer reviews" feststellbar.

Aus diesem Grund haben „peer reviews" eine überragende Bedeutung in der wissenschaftlichen Qualitätsbeurteilung. Sie werden angewandt bei der Entscheidung über Stellenbesetzungen, bei der Vergabe von Forschungsmitteln, bei Entscheidungen über Zeitschriften- oder Buchveröffentlichungen, bei der Evaluation ganzer Forschungseinrichtungen und bei Entwicklung von nationalen und internationalen Forschungseinrichtungen (Wouters et al., 2015: 44). Bei den heute besonders wichtigen Entscheidungen über Zeitschriften-Veröffentlichungen wird zumeist das doppelt-blinde Verfahren angewendet; [3] es gilt nachgerade als „heilige Kuh" (Osterloh & Kieser, 2015). [4]

[3] Bei diesem Verfahren, das insbesondere bei der Begutachtung von Zeitschriften-Artikeln zur Anwendung kommt, sollen (theoretisch) weder die Autoren die Gutachtenden noch umgekehrt die Gutachtenden die

### 1.2 Probleme der Peer Reviews

Allerdings gibt es eine Fülle von empirischer Evidenz dafür, dass die Gelehrtenrepublik mangelhaft funktioniert. *Erstens* gibt es eine geringe Übereinstimmung von Gutachterurteilen (Peters & Ceci, 1982; Bornmann & Daniel, 2003; Starbuck, 2005; Nicolai, Schmal & Schuster, 2015). Die Korrelation zwischen Gutachterurteilen liegt zwischen 0.09 und 0.5.[5] Dabei ist die Übereinstimmung von Gutachterurteilen im unteren Qualitätsbereich etwas höher als im oberen Bereich (Siler, Lee & Bero, 2015). In der klinischen Neurowissenschaft wurde sogar eine statistische Korrelation zwischen Gutachtern festgestellt, die nicht signifikant höher war als die einer Zufallsauswahl (Rothwell & Martyn, 2000). Die Auswahl und Anzahl der Gutachter hat einen entscheidenden Einfluss auf Annahme oder Ablehnung eines Papiers. Es ist dies das „luck of the reviewer draw"-Phänomen (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 (Starbuck, 2006).

*Drittens* ist die zeitliche Konsistenz von Gutachterurteilen niedrig. Es gibt zahlreiche Beispiele dafür, dass in sog. A-Journals zurückgewiesene Artikel später berühmt wurden und Preise gewonnen haben, inklusive des Nobel-Preises (Gans & Shepherd, 1994, Campanario, 1996; Siler, Lee & Bero, 2015). Ein aktuelles Beispiel ist Daniel Shechtman, der Chemie-Nobelpreisträger des Jahres 2011. Er wurde gemäss Zeitungsberichten[6] für seine Entdeckung der Quasi-Kristalle zunächst von seinen Kollegen nicht nur ausgelacht, sondern auch aus seiner Forschungsgruppe geworfen.

*Viertens* gibt es zahlreiche Bestätigungs-Fehler. Gutachter 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" liegend argumentiert (Mahoney, 1977).

---

Autorinnen kennen. Bei der Begutachtung von Forschungmitteln wird in der Regel das doppelt blinde Verfahren nicht angewendet. Hier kennen die Gutachtenden die Antragstellenden, nicht aber umgekehrt.

[4] Bei Entscheidungen über Forschungsmittel kommt zumeist das einfach-blinde Verfahren zur Anwendung. Bei Entscheidungen über Stellenbesetzungen wird mitunter auf die Anonymität der Gutachter verzichtet.

[5] Die Korrelation von Gutachterurteilen für den Schweizer Nationalfonds aus 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 favoritsm".

[6] Z.B. http://www.ftd.de/wissen/natur/:kopf-des-tages-daniel-shechtman-der-quasi-wissenschaftler/60112463.html

*Fünftens* gibt es einen beträchtlichen Institutionen- und Gender-Bias. Bei Forschungsanträgen favorisieren Gutachter Bewerbungen von prestigereichen Institutionen (Godlee et al., 1998). Der Nachweis eines Gender–Bias in Schweden bei der Vergabe von Forschungsgeldern hat viel Aufmerksamkeit erregt (Wenneras & Wold, 1999).

*Sechstens* erstellen beim doppelt-blinden Verfahren anonyme Gutachtende oft sehr oberflächliche Berichte. Ihre Kommentare sind mitunter wenig hilfreich, statt dessen setzen sie die Autoren erheblich unter Druck. Das Ergebnis ist „Publishing as Prostitution" (Frey, 2003).

*Siebtens* dauern Begutachtungsprozesse oft Monate, wenn nicht mehrere Jahre (Frey & Osterloh, 2014).

*Achtens* ist das System unverhältnismäßig teuer: Steuerzahlerinnen und –zahler werden von den Zeitschriften-Verlagen gleich fünffach zur Kasse gebeten: Zum ersten zahlt der Staat Saläre für die Verfasserinnen und Verfasser der Artikel, zum zweiten für die Gutachtenden und Editoren, 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 1000 US-Dollar oder oft auch mehr an die Verlage bezahlen.

Die Gelehrtenrepublik als Marktersatz funktioniert nach diesen Befunden mangelhaft. Dennoch ist sie unverzichtbar. „Peer review"- Verfahren (nicht notwendigerweise doppelt blinde Verfahren) bilden deshalb die Basis aller anderen Verfahren der wissenschaftliche Qualitätsbeurteilung. Zudem haben „peer reviews" einen wichtigen Vorteil, nämlich Vieldimensionalität, Dezentralität und Vielfalt. Wird eine Publikation abgelehnt, kann man sie in anderen Journals ähnlicher Qualität einreichen. Auch herrscht im deutschsprachigen System eine große Vielfalt an Möglichkeiten zur Bewerbung an gleichwertigen Universitäten.[7] Dies bringt aber ein Problem mit sich: Die Öffentlichkeit, d.h. Forschungsmanager, Journalisten und

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

Ministerien sowie Wissenschaftler anderer Fächer 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. Solche Kennzahlensysteme haben sich in den letzten Jahren fast flächendeckend etablieren können. Sogar Kollegen aus der Gelehrtenrepublik halten vielfach Kennzahlen für unverzichtbar, nämlich bei Anträgen für Forschungsmittel sowie als Grundlage für das „informed peer review" bei der Beurteilung von Bewerberinnen und Bewerbern. Dabei geht es um den Einbezug von Kennzahlen bei einer (ansonsten qualitativen) Qualitätsbeurteilung (Butler, 2007; Moed, 2007).

### 1.3 Probleme von Kennzahlen und Rankings

Als eine der wichtigsten Kennzahlen hat sich die Anzahl von doppelt blind begutachteten Artikeln etablieren können, die Forschenden in „guten" Journals , d.h. in Journals mit einem hohen Impact Faktor veröffentlichen.[8] 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, d.h. durch ein 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 (z.B. Archambault und Larivière, 2009; Jarwal, Brion & King, 2009). Etwas anders geht das Zeitschriftenranking des Verbandes der Hochschullehrer für Betriebswirtschaft (VHB-Jourqual) vor. Hier bewerten die Kolleginnen und Kollegen Journals nach ihrer Reputation. Auch hier wird unterstellt, dass die Qualität eines einzelnen Aufsatzes nach der Qualität der Zeitschrift bemessen werden kann, in welcher der Aufsatz veröffentlicht wurde. In beiden Fällen – Bewertung nach Impact-Faktor und nach Reputation - ist dies aber ein unsinniges Kriterium. Wie inzwischen hinlänglich kritisiert (z.B. Oswald, 2007; Baum, 2010; Kieser, 2012; Frey & Osterloh, 2013; Osterloh & Frey, 2015), kann aus dem Impact-Faktor oder

[8] Ein weiterer wichtiger Indikator ist der Hirsch-Index (vgl. zur Darstellung z.B. Helbing & Balietti 2011). Ein Forschender hat einen Hirsch-Index von h, wenn h seiner oder ihrer Publikationen mindestens h-mal zitiert wurden. Eine Person hat z.B. einen h-Index von 15, wenn 15 ihrer Publikationen wenigstens 15 mal zitiert werden.

der Reputation einer Zeitschrift kein Rückschluss auf die Qualität eines *einzelnen* Artikels gezogen werden, der in dieser Zeitschrift veröffentlicht wurde: In allen Journals werden einige wenige Aufsätze häufig zitiert; die allermeisten hingegen selten oder gar nie.[9] Wer auch nur eine Grundausbildung in Statistik genossen hat, weiß, dass bei einer stark schiefen Verteilung Durchschnittswerte keine Aussagekraft haben. Gleichwohl verwenden Wissenschaftler, die es eigentlich besser wissen müssten, diese Art der Qualitätsbewertung bei der Entscheidung über die Karrieren von Nachwuchskräften! Vielfach ist eine Habilitation weitgehend Formsache, wenn entsprechend diesen Kriterien genügend Publikationen in sog. A-Journals erreicht werden. Ganz ähnlich wird bei Berufungen auf Professuren vorgegangen. Einige Universitäten zahlen auch noch Geldbeträge für Publikationen in „guten" Journals.

Die Einsicht, dass die Veröffentlichung in einem „guten" Journal nicht gleichzusetzen ist mit einer „guten" Publikation, setzt sich langsam, aber stetig durch. Die International Mathematical Union IMU (2008) hat vorgerechnet, dass unter einer bestimmten Konstellation 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-Faktor. Man irrt somit in 62% der Fälle, wenn man sich nach dem Impact-Faktor richtet! Der Schweizerische Nationalfonds und einige Schweizer Universitäten haben 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 (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-Faktor oder Reputation der Zeitschrift. Vielmehr haben solche Kriterien weitere schwerwiegende negative Nebenwirkungen: Sie verursachen einen enormen Publikationsdruck, belasten das ohnehin überlastete Gutachtersystem, reduzieren die intrinsische Motivation der Forschenden und verursachen „Ranking Games" auf individueller wie auf institutioneller Ebene (Osterloh & Frey, 2015; Welpe et al., 2015).

[9] Selbstverständlich haben Artikel in einem A-Journal eine besonders hohe Chance zur Kenntnis genommen und zitiert zu werden. Deshalb müssten eigentlich die Zitationen von Autoren in einem B- und C-Journal höher und die von Autoren in einem A-Journal niedriger bewertet werden (vgl. Balaban 2012).

Kennzahlen und Rankings sind deshalb noch weniger als (doppelt blinde) „peer reviews“ als Marktersatz geeignet. Sie können erstens keineswegs die Fehler und Schwächen von Gutachten durch Aggregation der Urteile ausgleichen, weil der Prozess der Aggregation zahlreiche neue Fehler erzeugt. Zweitens bewirken sie – anders als vielfältig und mitunter widersprüchliche Gutachten – eine Hierarchisierung der Wissenschaft, welche den offenen Diskurs innerhalb der Gelehrtenrepublik behindert: Ein Argument oder ein empirisches Ergebnis in einer Zeitschrift mit einem hohen Impact-Faktor gilt als besser als vergleichbare Veröffentlichungen in weniger guten Journals (obwohl, wie oben dargelegt, der Impact Faktor ein sinnloses Kriterium für die Qualität ist). Eine Wissenschaftlerin, die viele Aufsätze in High-Impact-Journals veröffentlich hat, gilt als exzellent und ihre Meinung hat besonderes Gewicht. Rankings – wie z.B. das Handelsblattranking in der Ökonomik - stellen anhand dieser Kriterien eine Rangordnung innerhalb der Disziplin her. Dadurch wird die für den wissenschaftlichen Diskurs dringend erforderliche Vielfalt durch Einfalt ersetzt. Der „zwanglose Zwang des besseren Arguments“ (Habermas, 2009: 144) wird behindert, welcher ebenso wie der offene Widerspruch zur Grundlage der Gelehrtenrepublik gehört.

## 2 Alternativen zur herrschenden Qualitätsbeurteilung

### 2.1 Altmetrics

Altmetrics sind Indikatoren, die aus den sozialen Medien abgeleitet werden (als Übersicht vg. Wouters et al., 2015: 68 ff.), d.h. beruhen nicht auf Gutachter-Urteilen. Altmetrics veranlassen jedoch wie alle Indikatoren „Ranking Games“ (Osterloh & Frey, 2015) und sind einem „Performanz Paradoxon“ unterworfen (Meyer & Gupta, 1994; Meyer, 2009; Frost & Brockmann, 2015). Damit ist gemeint, daß alle Leistungsindikatoren mit der Zeit ihre Relevanz verlieren, sodass sie nicht mehr gute von schlechten Leistungen unterscheiden können. Die Ursache sind zwei gegenläufige Effekte, die allerdings in der Realität nur schlecht auseinander gehalten werden können: Leistungsindikatoren können einerseits einen positiven Lerneffekt hervorrufen, z.B. deutlich machen, dass für Wissenschaft Publikationen wichtig sind. Dies kann positive Anreiz-, Selektions- und Selbst-Selektions-Effekte bewirken, wodurch die Varianz der Leistung sinkt. Andererseits können sie auch einen perversen Lerneffekt erzeugen, der

durch „gaming the game“ oder Manipulation entsteht.[10] Dies 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). Schlimmer noch, durch das Abstellen auf Leistungsindikatoren kann die Leistung verschlechtert werden. Dies haben Brembs, Button & Munafo (2013) für die Bio-Medizin nachgewiesen. Sie zeigen, dass Forschungs-Ergebnisse um so unzuverlässiger sind, je höher der Impact-Faktor des Journals ist, in dem sie veröffentlicht wurden.

Die einzige Methode, einem solchen verhängnisvollen Paradox zu entrinnen, wäre die ständige Veränderungen und Anpassungen der Indikatoren durch die Fachleute eines Gebietes. Diese aber haben meist ein Interesse daran, das zu verhindern, haben sie doch meist gemäß der Indikatoren in ihre Karriere investiert und Einfluss errungen. Das erklärt die Hartnäckigkeit, mit der z.B. am Glauben festgehalten wird, dass eine Veröffentlichung in einem „high-impact-journal“ ein „guter“ Leistungsausweis sei.

Gibt es Alternativen der Qualitätsbeurteilung, die statt auf Indikatoren auf den Diskurs in der „Gelehrtenrepublik“ abstellen, aber zugleich in Rechnung stellen, dass Uneinigkeit in der Gelehrtenrepublik notorisch ist?

### 2.2. Offenes Post-Publication-Peer-Review-Verfahren

Das *offene Post-Publication-Peer-Review-Verfahren* (Kriegeskorte, 2012; Osterloh & Kieser, 2015; Osterloh & Frey, 2015a) ist eine solche Alternative. Es sieht widersprüchliche Gutachten nicht als Problem, sondern als ein Zeichen solider und produktiver Wissenschaft an. Kontroversen zwischen Gutachten oder Kommentaren bieten Anlass für die Fortentwicklung der Wissenschaft. Dies ist allerdings nur dann der Fall, wenn es einen offenen wissenschaftlichen Diskurs unter Begutachteten und Gutachtenden gibt. Das ist bei der derzeitigen ex ante Doppelt-Blind-Begutachtung nicht möglich. Gutachtende sind hier nicht Diskussionspartner, sondern Diktatoren.

Das neue Verfahren unterscheidet sich von herkömmlichen Verfahren in zweifacher Hinsicht. Erstens findet es *nach* der Veröffentlichung statt, d.h. verhindert nicht die Weitergabe von Forschungsergebnissen durch ex-ante-Begutachtung (die, wie oben dargelegt, in der Regel nur ein geringe prognostische Qualität aufweist). Zweitens

[10] Im betriebswirtschaftlichen Accounting haben Chen, Parsley& Yang (2014) gezeigt, dass Firmen durch Lobbying ihren Leistungsausweis “verbessern”, weil sie u.a. die massgeblichen Leistungskriterien beeinflussen. Zahlreich sind die Ansätze zum “Earnings Management”, als Überblick vgl. Pfaff & Ising (2010).

ist es offen, d.h. Gutachter bzw. Kommentatoren und Autoren sind nicht anonym.

Das Verfahren sieht vor, dass Forschende einen etablierten Kollegen oder eine Kollegin als „Editor“ anfragen, ob er oder sie Kommentare einholt. Diese würden auf einer gemeinsamen Plattform veröffentlicht werden. Der „Editor“ würde genannt werden und dadurch – ähnlich heutigen Zeitschriften-Editoren - Reputation gewinnen. Die Kommentare sollen in der Regel namentlich gekennzeichnet sein, damit sie zitiert werden können. So besteht ein Anreiz, gehaltvolle Beiträge zu schreiben, die als eigenständige, zitierfähige Publikation gelten können. Die Verfasser des ursprünglichen Artikels können auf derselben Plattform antworten. Sind die Kommentare oberflächlich oder gar feindselig (wie dies bei anonymen Gutachten mitunter der Fall ist), schädigt dies die Reputation des Gutachtenden. Kommt ein lebendiger Diskurs zustande, können nach einiger Zeit die Ergebnisse als „state of the art“ präsentiert werden.

Dieses neue System wäre nicht nur erheblich billiger und schneller als das derzeitige System. Es würde vor allem dem offenen, wissenschaftlichen Diskurs in der „Gelehrtenrepublik“ die Bedeutung geben, welche für wissenschaftlichen Fortschritt unabdingbar ist.

Allerdings hat das System auch Nachteile. Der wichtigste Nachteil besteht darin, dass „Matthäus-Effekte“ („Wer hat, dem wird gegeben“) auftreten können. Prominente Forscher werden mehr und möglicherweise interessantere Kommentare erhalten, als unbekannte Nachwuchs-Wissenschaftler. Wer einen bekannten „Editor“ findet, hat grössere Chancen, dass ein lebendiger Diskurs zustande kommt. „Old boys networks“ werden eine Rolle spielen, auch wenn immerhin eine größere Transparenz herrscht als bei den bisherigen anonymen Verfahren, bei denen Netzwerkeffekte nicht immer deutlich werden (Osterloh & Kieser, 2015).

## 3 Gezielte Einschränkung der Qualitätsbeurteilung

Wie dargelegt, verursachen alle Qualitätsbeurteilungs-Systeme in der Wissenschaft beträchtliche Probleme. Dies führt zur Überlegung, Ausmaß und Frequenz von Qualitätsbeurteilungen zu beschränken (Reichert 2013) und die „Evaluitis“ (Osterloh & Frey, 2007) gezielt abzubauen. Ziel ist, in Analogie zur statistischen Fehlertheorie, den Fehler erster Art - die Abweisung einer richtigen Forschungs-Hypothese – zu minimieren. Dieser Fehler ist im Vergleich zum Fehler zweiter Art – der Annahme einer falschen Forschungs-Hypothese - gravierender. Der Fehler erster Art

führt dazu, dass richtige und erfolgversprechende Forschung verhindert wird. Dagegen bewirkt der Fehler zweiter Art lediglich, dass unnötige Forschung ermöglicht wird (Gillies, 2005). Dies bedeutet zwar eine Verschwendung von Forschungsmitteln, aber nicht die Verhinderung erfolgversprechender Forschung. Evaluationen konzentrieren sich zumeist auf den Fehler zweiter Art. Fehler erster Art können nur dadurch vermieden werden, dass erstens mehr Freiräume geschaffen werden, welche der Evaluation partiell entzogen werden. Zweitens muss angesichts des hohen Risikos fehlerhafter Beurteilung die Diversität wissenschaftlicher Leistung verstärkt werden. Damit werden - analog zur Evolutions- und Portfolio-Theorie – Risiken diversifiziert und die Chancen für innovative und überlebensfähige Ideen erhöht. Hierzu seien zwei Verfahren diskutiert.

### 3.1 Gezielte Einschränkung der Qualitätsbeurteilung durch Eingangskontrolle

Der *erste Vorschlag* will die Anlässe für Evaluationen auf wenige karrierererelevante Entscheidungen reduzieren, z.B. bei der Bewerbung um eine Stelle oder bei der Beantragung von Forschungsmitteln. Dieses Konzept hilft, die geschilderten Schwächen der Begutachtungsprozesse zu reduzieren, weil Begutachtungen auf wenige Anlässe beschränkt werden. Eine sorgfältige Eingangskontrolle ersetzt die kontinuierliche Bewertung durch dauernde Evaluationen (Osterloh, 2010; Frey & Osterloh, 2012; Osterloh & Frey, 2015). 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 Eingangs-Kontrolle ist keineswegs neu. Sie wird an den „Institutes for Advanced Studies“ ebenso praktiziert wie im “MacArthur Fellows Program” (Ioannidis, 2011) und an der Harvard-Universität. In deren Prinzipien heißt es: „The primary means for controlling the quality of scholarly activities of this faculty is through the rigorous academic standards applied in selecting its members.“[11] Das Konzept ist aber gleichwohl auf Gutachten mit allen den geschilderten Problemen angewiesen. Insbesondere bleiben die Probleme der „old-boys-networks“ bzw. des Favoritismus bestehen.

[11] http://www.fas.harvard.edu/research/greybook/principles.html.

### 3.2 Gezielte Einschränkung der Qualitätsbeurteilung durch partielle Zufallsauswahl

Der *zweite Vorschlag* ist radikaler. Er sieht die partiell zufällige Auswahl von Personen oder Forschungsprogrammen vor. Es ist die einzige Alternative, die – wenn die Zufallsauswahl kontrolliert und korrekt durchgeführt wird – Favoritismus und Manipulation vermeidet und zugleich die Diversität von Ideen gewährleistet.

Unbeabsichtigt bestimmt Zufall heute schon wesentliche Teile der Wissenschaft, jedoch weder kontrolliert noch korrekt. Die Auswahl von Artikeln durch Gutachten grenzt heute schon in einigen Fällen an Zufall (Bornmann & Daniel, 2009; Siler, Lee & Bero, 2015; Nicolai, Schmal & Schuster, 2015). Nicht weniger als ein Drittel von Forschungsmitteln wurde 2009 vom Australischen National Health and Medical Research Council aufgrund Probleme des „peer review" Systems faktisch zufällig ausgewählt (Graves, Barnett & Clarke, 2011).

Zufall in absichtlicher und kontrollierter Form hat in der Vergangenheit in zahlreichen Verfahren ein Rolle gespielt (Buchstein, 2009). Im klassischen Athen und im mittelalterlichen Venedig wurden politische Positionen in einem gemischten Verfahren aus Zufall und gezielter Auswahl besetzt. Auch andere italienische Stadtstaaten des Mittelalters wie Florenz haben Elemente des Zufallsverfahrens zur Bestimmung ihrer Exekutive verwendet (Dumler, 2001). An der Universität Basel wurden im 18. Jahrhundert Lehrstühle per Zufallsauswahl aus einer Liste von drei Kandiaten ausgewählt (Burckhardt, 1916; Stolz, 1986). Heute noch wird der koptische Papst aus drei zuvor ausgewählten Personen bestimmt (Boochs, 2009). In der Literatur zur deliberativen Demokratie wird vorgeschlagen, zufällig ausgewählte Bürger im Entscheidungsprozess zu beteiligen (Dryzek, 2000; Habermas, 2006). Zeitoun, Osterloh & Frey (2014) haben Zufallsverfahren für die Corporate Goverance erörtert.

„Zufall" wird hier im Sinne einer statistischen *Wahrscheinlichkeit* verwendet. Es hat somit nichts mit Willkür zu tun, sondern eher mit dem Gegenteil, einer strengen mathematischen Gesetzmäßigkeit. Wie alle Entscheidungsverfahren haben Zufallsprozeduren Vor- und Nachteile (Buchstein 2009, Zeitoun, Osterloh & Frey, 2015; Frey & Steiner, 2014). Die wichtigsten Vorteile sind die folgenden:

*Erstens* schützen Zufallsentscheidungen vor Favoritismus und „old boys networks". Es lohnt sich nicht, vor der Wahl in Lobbying, Bestechung oder andere

Einflussversuche zu investieren, wenn zufällig entschieden wird. Damit werden auch die Kosten für eine Kandidatur, z.B. für (Selbst-)Marketing obsolet, was zu einem größeren Kandidaten-Pool führt.

*Zweitens* werden Kandidierende ermutigt, die sonst nicht in Betracht gezogen, übersehen oder marginalisiert würden. Zufallsauswahl ist deshalb eine „Suchmaschine" für neue Perspektiven und Talente (Buchstein, 2009: 391) sowie für solche Personen, die das Risiko der Ablehnung, einen damit verbundenen möglichen Reputatiosverlust oder den Wettbewerb scheuen. Das ist – so zahlreiche empirische Befunde – vor allem bei Frauen der Fall (Gneezy et al., 2003; Niederle & Vesterlund, 2007; Niederle, Segal & Vesterlund, 2013; Balafoutas & Sutter, 2012). Zufallsauswahl ist deshalb besonders geeignet, mehr Frauen zu veranlassen, sich als Kandidatinnen zur Verfügung zu stellen (Goodall & Osterloh, 2015).

*Drittens* fördert Zufall neue Ideen zutage, die im herkömmlichen Betrieb wenige Chancen haben. Häufig sind es die Ideen „von aussen" welche die Kreativität dank einer nützlichen Ignoranz des „herrschenden Wissens" oder einer "focused naïveté" (Gieryn & Hirsh, 1984: 91) beflügeln. Dies zeigen empirische Befunde zur Innovationsforschung (Jeppesen & Lakhani, 2010; Rost & Osterloh, 2010; Talke, Salomon & Rost, 2010)

*Viertens* verlieren bei der Zufallsauswahl die Verlierer der Wahl nicht das Gesicht und ihr Selbstwertgefühl, wie das bei normalen Auswahlprozessen oft der Fall ist. Dies führt ebenfalls zu einer Vergrößerung des Kandidierenden-Pools. Dieser Sachverhalt spielte z.B. bei der Praktizierung der Zufallsauswahl an der Universiät Basel im 18. Jahrhundert eine entscheidende Rolle (Burckhardt, 1916).

*Fünftens* führt Zufall zu einer präzisen Repräsentativität der Grundgesamtheit. Auch dies führt dazu, dass Personenkreise zum Zuge kommen, die sonst leicht übersehen werden. Gegenüber Quoten besteht der große Vorteil, dass dort von vorneherein festgelegt werden muss, welche Dimensionen (etwa Geschlecht, Alter, Nationalität) als relevant angesehen werden. Bei Zufallsverfahren ist dies nicht der Fall. Damit können bisher vernachlässigte Perspektiven Beachtung finden (Frey & Steiner, 2014) .

*Sechstens* erleichtert das Zufallsprinzip Stabilität und Kontinuität, wenn es Gruppen mit divergierenden Interessen gibt. Jede dieser Gruppen hat die Chance, zum Zuge zu kommen, selbst wenn bisher die Gegenpartei dominierte. Dieser Aspekt spielte im klassischen Athen und in den italienischen Stadtstaaten des Mittelalters eine grosse

Rolle. Deren Gedeihen war immer wieder durch politische Unruhen und Bürgerkriege gefährdet (Duxbury, 1999; Stone, 2009).

Diesen Vorteilen stehen auch Nachteile gegenüber. Der *erste* und wichtigste Nachteil besteht darin, dass nicht zwischen fähigen und unfähigen Personen unterschieden wird. Aus diesem Grund gibt es selten reine Zufallsverfahren, sondern diese werden fast immer mit einer Auswahl aus einer Grundgesamtheit gemäß ihren Qualitäten kombiniert.

Ein *zweiter* Nachteil besteht darin, dass Zufallsauswahl manchmal als "irratonal" angesehen wird. Allerdings bergen intendiert rationale Entscheidungsprozesse oft genug faktische Irrationalitäten in sich. Bereits hingewiesen wurde darauf, dass die Auswahl von Journal-Artikeln oder die Zuweisung von Forschungsmitteln faktisch oft an Zufall grenzt. Auch sonst gibt es zahlreiche Verzerrungen sogenannt „rationaler" Entscheidungsprozesse (Kahnemann, 2011). Darüber hinaus hat sich gezeigt, dass die Preisverleihung bei Musik-Wettbewerben (Ginsburgh & Weyers, 2014) oder die Prämierung von Wein (Hodgson, 2009) faktisch zufällig ist. Erinnert sei auch an das „Peter Prinzip" bei der Beförderung innerhalb von Hierarchien (Laurence & Hull, 2001) und an die oben angeführten Überlegungen zum „Performanz Paradoxon". Faktisch ist also die Rationalität von Auswahl- und Entscheidungsprozessen oft nur Fassade, die oft genug als unfair empfunden wird. Eine an mathematischen Wahrscheinlichkeiten ausgerichtete Zufallsauswahl erscheint demzufolge im Vergleich dazu als durchaus rational.

Die partielle Anwendung von Zufallsverfahren in der Wissenschaft wird dadurch erleichtert, daß gemäß empirischen Befunden Gutachterurteile verlässlicher sind, soweit es sich um schlechte Beiträge handelt (Cicchetti, 1991; Moed, 2007; Siler, Lee & Bero, 2015). Es besteht mehr Einigkeit unter den Gutachtenden, welche Beiträge abzulehnen sind. Bei guten oder mittleren Beiträgen unterscheiden sich die Einschätzungen hingegen oft diametral. Man kann deshalb davon Gebrauch machen, dass mittels Gutachten eine Vorauswahl getroffen wird, welche Kandidierenden oder Anträge auf Forschungsmittel nicht in Frage kommen. Genauso kann man in den (meist selteneren) Fällen verfahren, in denen unter allen Gutachtenden Übereinstimmung im positiven Sinne herrscht. Die Zufallsauswahl könnte dann in den (meist überwiegenden) Fällen zur Anwendung kommen, bei denen Dissens herrscht. Es ist zu vermuten, dass sich hierbei nicht selten um neuartige und ungewöhnliche Personen oder Beiträge

handelt, die ansonsten wenig Chancen haben, sich im etablierten Wissenschaftsbetrieb durchzusetzen.

Dieser Vorschlag findet vermutlich für die Auswahl von Forschungsprojekten in der Forschungsförderungspolitik leichter Akzeptanz als bei der Auswahl von Personen oder von Veröffentlichungen. Die Kombination von Experten-Urteilen und Zufall könnte deshalb bei Forschungsmitteln – zumindest in einem zeitlich begrenzten Experiment - für einen bestimmten Prozentsatz der Forschungsmittel angewendet werden (Buchstein, 2011). Nach einigen Jahren ließe sich empirisch ermitteln, ob die partiell zufällig ausgewählten Forschungsprojekte den wissenschaftlichen Diskurs besser oder schlechter befruchtet haben als die nach konventionellen Verfahren ausgewählten Projekte.

Gegen die vorgeschlagenen neuen Verfahren werden diejenigen Protest anmelden, welche mit Hilfe des gegenwärtigen Systems Gewinne erzielt und Einfluss errungen haben. Auch werden Übergangs-Probleme auftreten. Aber die Probleme des heutigen wissenschaftlichen Qualitätsbeurteilungs-Systems sind so riesig (vgl. The Economist, 2013), dass dringend Alternativen aufgezeigt, ernsthaft diskutiert und ausprobiert werden sollten.

***Referenzen***

Alberts, B. (2013). Editorial: Impact Factor Distortions. *Science*, 17. May 2013: 787.

Archambault, É. & Larivière, V. (2009). History of the journal impact factor: contingencies and consequences. *Scientometrics*, 79(3): 639-653.

Balafoutas, L. & Sutter, M. (2012). Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science***,** 335(6068): 579-582.

Balanban, A. T. (2012). Positive and negative aspects of citation indices and journal impact factors. *Scientometrics*, 92: 241-247.

Baum, J. A. C. (2010). Free-Riding on Power Laws: questioning the validity of the Impact Factor as a measure of research quality in organization studies. *Organization*, 18 (4): 449-466.

Boochs, W. (2009). *Geschichte und Geist der Koptischen Kirche.* 2 ed. Aachen: Bernardus-Verlag.

Bornmann, L. (2011). Scientific Peer Review. *Annual Review of Information Science and Technology*: 199-245.

Bornmann, L. & Daniel, H. D. (2003). Begutachtung durch Fachkollegen in der Wissenschaft. Stand der Forschung zur Reliabliliät, Fairness und Validität des Peer-Review-Verfahrens. In: Schwarz, S. & Teichler, U. (Hrsg.). *Universität auf dem Prüfstand. Konzepte und Befunde der Hochschulforschung.* Frankfurt a.M., Campus: 211-230.

Bornmann, L. & Daniel, H. D. (2009). The luck of the referee draw: The effect of exchanging reviews. *Learned Publishing*, vol. 22(2): pp. 117–25.

Buchstein, H. (2009). *Demokratie und Lotterie: Das Los als politisches Entscheidungsinstrument von der Antike bis zur EU*: Campus Verlag.

Buchstein, H. (2011). Der Zufall in der Forschungsförderungspolitik. *Forschung & Lehre* 11(8): 596–597.

Butler, L. 2007. Assessing University Research: A Plea for a Balanced Approach. *Science and Public Policy* 34: 565–74.

Brembs, B., Button, K. & Munafo, M. (2013). Deep impact: unintended consequences of journal rank. *Frontiers in Human Neuroscience* 7: 1-11. doi: 10.3389/fnhum.2013.00291

Burckhardt, A. (1916). Ueber die Wahlart der Basler Professoren, besonders im 18. Jahrhundert. *Basler Zeitschrift für Geschichte und Altertumskunde* (15): 28-46.

Campanario, J. M. (1996). Using citation classics to study the incidence of serendipity in scientific discovery. *Scientometrics*, vol. 37: 3–24.

Chen, H., Parsley, D. C. & Yang Y.-W. Corporate Lobbying and Financial Performance http://ssrn.com/abstract=1014264.

Cicchetti, D. V. (1991). The Reliability of Peer Review for Manuscript and Grant Submissions: A Cross-disciplinary Investigation. *Behavioral and Brain Sciences* 14: 119–35.

Contopoulos-Ioannidis, D. G., Alexiou, G. A., Gouvias, Th.c. & Ioannidis J. P. A. (2008). Life Cycle of Translational Research for Medical Interventions. *Science*, 321, September 2008: 1298-1299.

Davies, P. (1995). *Die Unsterblichkeit der Zeit. Die moderne Physik zwischen Rationalität und Gott.* Scherz, Bern u.a.

DORA (San Francisco Declaration on Research Assessment) (2012). Accessed November 18, 2014. http://am.ascb.org/dora/files/SFDeclarationFINAL.pdf.

Dryzek, J. S. (2000). *Deliberative democracy and beyond: Liberals, critics, contestation*. Oxford, UK: Oxford University Press.

Dumler, Helmut (2001). *Venedig und die Dogen*. Düsseldorf: Artemis & Winkler.

Duxbury, Neil (1999). *Random justice: on lotteries and legal decision-making*: Clarendon Press.

Frey, B. S. (2003). Publishing as prostitution? – Choosing between one's own ideas and academic success. *Public Choice*, 116: 205–223.

Frey, B. S. & Pommerehne, W. (1990). On the fairness of pricing – An empirical survey among the gerneral population. *Journal of Economic Behaviour and Organization*, 20: 295-307.

Frey, B. S. & Kirchgässner G. (1994). *Demokratische Wirtschaftspolitik.* 2. Aufl. München: Vahlen.

Frey, B. S. & Osterloh, M. (2012). Rankings: Unbeabsichtigte Nebenwirkungen und Alternativen. *Ökonomenstimme* 17. Februar 2012.

Frey, B. S. & Osterloh, M. (2013). Gut publizieren = gute Publikation? *Ökonomenstimme* 16. Mai 2013.

Frey, B. S. & Osterloh, M. (2014). Schlechte Behandlung des wissenschaftlichen Nachwuchses und wie man das ändern könnte. *Ökonomenstimme* 28. Oktober 2014.

Frey, B. S. & Osterloh, M. (2016). Impact Faktoren: Absurde Vermessung der Wissenschaft. *Soziale Welt* (im Druck).

Frey, B. S & Steiner, L. (2014). Zufall als gesellschaftliches Entscheidungsverfahren. In: *Recht im ökonomischen Kontext - Festschrift zu Ehren von Christian Kirchner.* Wulf A. Kaal, Matthias Schmidt und Andreas Schwartze (Hg.). Mohr Siebeck, Tübingen, 2014: 749 – 761

Frost, J. & Brockmann, J. (2015). When quality is equated with quantitative productivity – Scholars caught in a performance paradox. *Zeitschrift für Erziehungswissenschaft* (in press).

Gans, J. S. & Shepherd, G. B. (1994). How are the mighty fallen: Rejected classic articles by leading economists. *Journal of Economic Perspectives*, vol. 8: 165–79.

Gillies, D. (2005). *Lessons from the History and Philosophy of Science regarding the Research Assessment Exercise.* Paper read at the Royal Institute of Philosophy on 18. November 2005. (www.ucl.ac.uk/sts/gillies).

Gieryn, T. F. & Hirsh, R. F. (1984). Marginalia: Reply to Simonton and Handberg. *Social Studies of Science,* 14(4): 624-624.

Ginsburgh, V. & Weyers, S. (2014). Nominees, winners, and losers. *Journal of Cultural Economics*, 38: 291-313.

Gneezy, U., Niederle, M. & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics*, 118(3): 1049-1074.

Godlee, F., Gale, C. 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. *Journal of the American Medical Association*, 263, 10: 1438-1441.

Goodall, A. & Osterloh, M. (2015). Room at the top. *Times Higher Education* 14. May 2015: 35-39.

Graves, N., Barnett, A. G. & Clarke, P. (2011). Cutting random funding decisions. *Nature* (469): 299.

Habermas, J. (2006). Political communication in media society. *Communication Theory*, 16: 411–426.

Habermas, J. (2009). Wahrheitstheorien. Philosophische Texte. Studienausgabe in fünf Bänden, Frankfurt am Main 2009, Bd. 1.

Hattrup, D. (2011). *Einstein gegen den würfelnden Gott*. Kindle Edition, 2011.

Helbing, D. & Balietti, S. (2011). How to create an innovation accelerator. *EPJ Special Topics* 195: 101-136.

Hodgson, R. (2009). An analysis of the concordance among 13 wine competitions. *Journal of Wine Economics*, 4: 1–9.

International Mathematical Union IMU (2008). *Citation Statistics*. A report. Corrected version, 16/12/08.

Jeppesen, L. B. & Lakhani, K. R. (2010). Marginality and problem-solving effectiveness in broadcast search. *Organization Science*, 21(5): 1016-1033.

Ioannidis, J. P. A. (2011). More time for research: Fund people not projects. *Nature* (477): 529-531,

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? *Journal of Higher Education Policy and Management*, 31(4): 289-300.

Kahnemann, D. (2011). *Schnelles Denken, langsames Denken*. München: Siedler.

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

Kriegeskorte N. (2012). Open evaluation: a vision for entirely transparent post-publication peer review and rating for science. *Frontiers in Computational Neuroscience* 6: 1–18.

Laband, D. N. (2013). On the Use and Abuse of Economics Journal Rankings. *The Economic Journal*, vol.123: F223-54.

Laurence, J. P. & Hull, R. (2001). Das Peter-Prinzip oder die Hierarchie der Unfähigen. Übersetzt von Michael Jungblut, 12. Auflage, Rowohlt-TB 61351, Reinbek bei Hamburg.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy Research*, vol. 1: 161–75.

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

Merton, R. K. (1973). The sociology of science: Theoretical and empirical investigation, Chicago, IL: University of Chicago Press.

Meyer, M. W. & Gupta, V. (1994). The Performance Paradox. *Research in Organizational Behavior* 16: 309-369.

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

Moed, H. F. (2007). The Future of Research Evaluation Rests with an Intelligent Combination of Advanced Metrics and Transparent Peer Review. *Science and Public Policy* 34: 575–83.

Nicolai, A. T., Schmal, S. & Schuster, Ch. (2015). Interrater Reliability of the Peer Review Process in Management Journals. In: I. Welpe, J. Wollersheim, S. Ringelhan & M. Osterloh (Hrsg.), *Incentives and Performance - Governance of Research Organization*: Springer Verlag, 2015, Heidelberg: 107-120.

Niederle, M. & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics*, 122(3): 1067-1101.

Niederle, M., Segal, C. & Vesterlund, L. (2013). How costly is diversity? Affirmative action in light of gender differences in competitiveness. *Management Science*, 59(1): 1-16.

Osterloh, M. (2015). Die Gelehrtenrepublik funktioniert nur mangelhaft. *Akademie Aktuell* (3): 24-29.

Osterloh, M. (2010). Governance by Numbers. Does it really work in Research? *Analyse und Kritik*, Vol. 32(2): 267–83.

Osterloh, M. & Frey, B. S. (2007). Die Krankheit der Wissenschaft. *Frankfurter Allgemeine Zeitung* 21.7.2007 (167): 13.

Osterloh, M. & Frey, B. S. (2015). Ranking Games. *Evaluation Review.* Vol 39(1): 102-129. DOI: 10.1177/0193841X14524957.

Osterloh, M. & Frey B. S. (2015a). Ranking Games und wie man sie überwinden kann. *Zeitschrift für Kulturwissenschaft*, (1): 65-80.

Osterloh, M. & Kieser, A. (2015). Double-Blind Peer Review: How to Slaughter a Sacred Cow. In: I. Welpe, J. Wollersheim, S. Ringelhan & M. Osterloh (Hrsg.), Incentives and Performance - Governance of Research Organizations (307–324): Springer International Publishing, Cham et al.

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

Peters, D. & Ceci, S. J. (1982). Peer review practices of psychological journals: The fate of published articles, submitted again. The Behavioral and Brain Sciences, 5: 187–195.

Peters, K., Daniels, K., Hodgkinson, G. P. & Haslam, A. (2014). Experts´ Judgements of Management Journal quality: In Identity Concerns Model. *Journal of Management*, 40: 1785-1812.

Pfaff, D. & Ising, P. (2010). Earnings Management. Erscheinungsformen und Aufdeckungsmöglichkeiten, in: Seicht, G. Jahrbuch für Controlling und Rechnungswesen. Wien: 291-312.

Polanyi, M. (1962/2002). The republic of science: Its political and economic theory. Minerva, vol. 1, pp. 54–73. Reprinted in Polanyi, M. (1969). From Knowing and Being, pp. 49–72, Chicago: University of Chicago Press. Re-reprinted in Mirowski, P. and Sent, E.M. (2002) (eds). Science Bought and Sold. *Essays in the Economics of Science*, pp. 465–85, Chicago: The University of Chicago Press.

Reichert, S. (2013). Jenseits der Leistungsüberprüfung - Diskussionspapier zur Suche nach einem neuen Umgang mit Qualitätssicherung an Hochschulen. Diskussionspapier des Schweizerischen Wissenschafts- und Technologierates (SWTR).

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

Rost, K. & Osterloh, M. (2010). Opening the Black Box of Upper Echelons: Expertise and Gender as Drivers of Poor Information Processing. Corporate Governance. *An International Review*, 2010, Vol. 18(3): 212-233.

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? *Brain*, vol. 123: 1964–9.

Siler, K., Lee, K. & Bero, L. (2015). Measuring the effectiveness of scientific gatekeeping. *PNAS*, 112 (2): 360-365.

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

Smith, Richard (2015). Ineffective at any dose? Why peer review simply doesn't work. *Times Higher Education*, 28 May 2015, Opinion.

Starbuck, W. H. (2005). How much better are the most prestigious journals? The statistics of academic publication. *Organization Science*, 16: 180-200.

Starbuck, W. H. (2006). The production of knowledge. The challenge of social science research, Oxford: Oxford University Press.

Stolz, P. (1986). Parteienwettbewerb, politisches Kartell und Tausch zwischen sozioökonomischen Gruppen. *Schweizerische Zeitschrift für Volkswirtschaft und Statistik*, 122: 657–675.

Stone, P. (2009). The Logic of Random Selection. *Political Theory*, 37: 375-397.

Strathern, M. (1996). From Improvement to Enhancement: An Anthropological Comment on the Audit Culture. *Cambridge Anthropology* 19:1–21.

Talke, K., Salomo, S. & Rost, K. (2010). How top management team diversity affects innovativeness and performance via the strategic choice to focus on innovation fields. *Research Policy*, Vol. 39(7): 907-918.

The Economist (2013). How science goes wrong. http://www.economist.com/news/leaders/21588069.

Welpe, J., Wollersheim, J., Ringelhan, S. & Osterloh, M. (2015). Preface. In: Dies. (Hrsg): *Incentives and Performance - Governance of Research Organization*: Springer Verlag, 2015, Heidelberg: v - xxii.

Wenneras, C. & Wold, A. (1999). Bias in peer review of research proposals in peer reviews in health sciences. In: F. Godlee & T. Jefferson (Eds.), *Peer review in health sciences* (pp. 79–89). London, UK: BMJ Books.

Wouters, P. et al. (2015). *The Metric Tide. Literature Review. Supplementary Report I to the Independent Review of the Role of Metrics in Research Assessment and Management.* HEFCE. DOI:10.13140/RG.2.1.5066.3520

Zeitoun, H., Osterloh, M. & Frey, B. S. (2014). Learning from Ancient Athens: Demarchy and corporate governance. *Academy of Management Perspectives*, 28(1): 1–14.