Begutachtungswesen

Die Gelehrtenrepublik funktioniert nur mangelhaft

Systematisches Marktversagen, Impact-Faktoren und Rankings: über die Evaluation des wissenschaftlichen Nachwuchses.

Von Margit Osterloh

Warum ist die Evaluation des wissenschaftlichen Nachwuchses so schwierig?

Es gibt wohl wenige Berufsgruppen, die so häufig und so intensiv evaluiert werden wie Nachwuchswissenschaftlerinnen und -wissenschaftler: Haben sie ihre Diplom- oder Masterprüfung abgeschlossen, müssen sie sich bei der Aufnahme in das Promotionsstudium bewerben, Anträge auf Fördermittel schreiben, sich mit ihren Papieren bei Konferenzen bewerben, ihre Aufsätze bei wissenschaftlichen Zeitschriften begutachten lassen, nach Ablehnungen wieder und wieder einreichen, zumeist jährli-

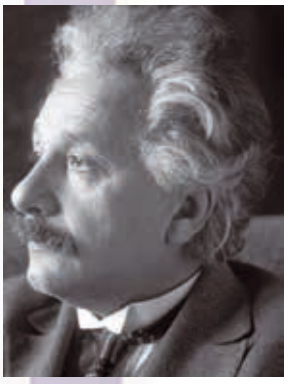
ABB.: CHRISTIAN MURZ / FOTOCOMMUNITY; WIKIMEDIA COMMONS (4)

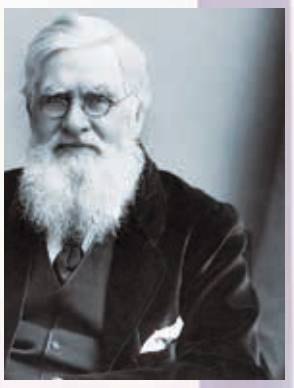
che Evaluationen ihrer heimischen Forschungsinstitution über sich ergehen lassen, ihre Dissertation begutachten lassen, neue Aufsätze bei wissenschaftlichen Zeitschriften einreichen, die Habilitation begutachten lassen, Bewerbungen für Professuren einreichen, wiederum Anträge für Drittmittel stellen, immer wieder Aufsätze begutachten lassen und so weiter und so weiter. Dabei stehen die Forschenden auch noch unter einem enormen Zeitdruck, weil sie bis zur Professur meist nur befristete Anstellungsverträge haben.

Systematisches Marktversagen

Das umständliche Evaluationssystem ist der Tatsache zu verdanken, dass es in der Wissenschaft ein systematisches Marktversagen gibt. Dieses entsteht einerseits dadurch, dass Wissenschaft öffentliche Güter produziert, die durch Nichtausschließbarkeit bei der Nutzung und Nichtrivalität im Konsum des produzierten Wissens gekennzeichnet sind.

Zum Zweiten ist Forschung gekennzeichnet durch fundamentale Unsicherheit. Diese ist sichtbar an so genannten Serendipitätseffekten: Man findet etwas anderes als das, was man gesucht hat. Solche Effekte sind in der Wissenschaft zahlreich, wie man etwa an der Entdeckung des Dynamits, der Röntgenstrahlen oder der Radioaktivität sehen kann. Drittens stellt sich der Nutzen wissenschaftlicher Entdeckungen mitunter erst nach sehr langer Zeit ein. In der Wissenschaft handelt es sich daher um Vertrauensgüter im Unterschied zu Erfahrungsgütern. Bei Letzteren kann man nach Gebrauch feststellen, ob sie etwas taugen oder nicht. Bei Vertrauensgütern kann man das nur sehr langfristig oder manchmal nie. Zum Vierten gibt es Schwierigkeiten, einzelne Entdeckungen bestimmten Personen zuzurechnen. Die Wissenschaftsgeschichte ist voll von so genannten Multiples, also Entdeckungen, die ursprünglich Einzelnen zugeschrieben wurden und die sich später als „in der Luft liegend" herausgestellt haben. Hier ist also nicht klar, wer der Entdecker war. Dazu gehört beispielsweise die Erfindung der Infinitesimalrechnung, bei der nicht klar ist, ob sie zuerst auf Leibniz oder Newton zurückgeht. Ebenso ist es mit der Allgemeinen Relativitätstheorie, bei der ex post Zweifel existieren, ob Einstein oder Hilbert der Entdecker war, oder mit der Evolutionstheorie mit Darwin versus Wallace.





Die Wissenschaftsgeschichte ist voll von so genannten Multiples: David Hilbert versus Albert Einstein (o.), Alfred Russel Wallace versus Charles Darwin.

Die Gelehrtenrepublik als Marktersatz

Für den fehlenden Markt braucht Wissenschaft einen Ersatz. Das ist die Gelehrtenrepublik, die „Republic of Science". Diese stellt mit Gutachten fest, wer gute Forschung macht. Das bringt eine Formulierung des Philosophen Michael Polanyi zum Ausdruck: „The soil of academic science must be exterritorial in order to secure its rule by scientific opinion." Die Qualität der Forschenden ist also nur von „innen heraus" durch die „Scientific Community" feststellbar.

Leider gibt es eine Fülle von empirischer Evidenz dafür, dass die Gelehrtenrepublik mangelhaft funktioniert. Erstens belegt eine Reihe von Untersuchungen, dass Gutachterurteile nur in geringem Maße übereinstimmen. Die Korrelation zwischen Gutachterurteilen liegt zwischen 0.09 und 0.5. Dabei ist die Übereinstimmung von Gutachterurteilen im unteren Qualitätsbereich höher als im oberen Bereich. In der klinischen Neurowissenschaft wurde sogar eine statistische Korrelation zwischen Gutachtern festgestellt, die nicht signifikant höher war als die einer Zufallsauswahl. Die Auswahl der Gutachter hat einen entscheidenden Einfluss auf Annahme oder Ablehnung eines Papiers. Zweitens ist die prognostische Qualität von Gutachten gering. Die Reviewer-Einschätzungen korrelieren nur mit 0.25 bis 0.37 mit späteren Zitationen.

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 Nobelpreises. Ein aktuelles Beispiel ist Daniel Shechtman, der Chemie-Nobelpreisträger des Jahres 2011. Er wurde gemäss Zeitungsberichten für seine Entdeckung der Quasikristalle zunächst von seinen Kollegen nicht nur ausgelacht, sondern auch aus seiner Forschungsgruppe hinausgeworfen. Viertens gibt es zahlreiche Bestätigungsfehler: Gutachter fanden in 72 Prozent von Papieren methodische Fehler, wenn diese dem „Mainstream" widersprachen, hingegen nur in 25 Prozent der Fälle, wenn das Papier im „Mainstream" liegend argumentierte. Fünftens gibt es einen beträchtlichen Institutionen- und Gender-Bias. Bei Forschungsanträgen favorisieren Gutachter Bewerbungen von prestigereichen Institutionen. So hat etwa der Nachweis eines Gender-Bias in Schweden bei der Vergabe von Forschungsgeldern vor einigen Jahren viel Aufmerksamkeit erregt.

Impact-Faktoren und Rankings als Evaluationsunterstützung?

Die Gelehrtenrepublik als Marktersatz funktioniert nach diesen Befunden schlecht, obwohl sie auch Vorteile hat, nämlich Vieldimensionalität, Dezentralität und Vielfalt. Wird eine Publikation abgelehnt, kann man sie in anderen Journals ähnlicher Qualität einreichen. Auch herrschen im deutschsprachigen Universitätssystem zahlreiche Möglichkeiten, sich an gleichwertigen Universitäten zu bewerben. Dies bringt aber ein Problem mit sich: Die Öf-

fentlichkeit, d. h. Forschungsmanager, Journalisten 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.

Als solche Kennzahlen haben sich die Anzahl von Artikeln etablieren können, die Forschende in „guten" Journals (A-Journals) veröffentlichen, sowie die sich daraus ergebenden Rankings. Dabei wird unterstellt, dass ein in einer „guten Zeitschrift" veröffentlichter Artikel auch eine „gute Publikation" darstellt, weil solche Zeitschriften die „kollektive Weisheit" einer „Scientific Community" darstellen. Was eine „gute" Zeitschrift ist, wird meist durch den Impact-Faktor 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 heute international durchgesetzt. Etwas anders geht das VHB-Jourqual vor, das Zeitschriftenranking des Verbandes der Hochschullehrer für Betriebswirtschaft. 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, kann aus dem Impact-Faktor 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 nie. 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 A-Journals erreicht werden. Ganz ähnlich wird bei Berufungen auf Professuren vorgegangen. Einige Universitäten zahlen auch noch Geldbeträge für Publikationen in „guten" Journals. Dabei ist es eine Selbstverständlichkeit, dass Artikel in einem A-Journal eine besonders hohe Chance haben, 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.

A- und C-Journale: Wie sinnvoll ist es, einen einzelnen Aufsatz nach der Qualität der Zeitschrift zu bewerten, in der er erscheint?

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) hat vorgerechnet, dass die Wahrscheinlichkeit, dass ein zufällig ausgewählter Artikel in einer Zeitschrift mit einem niedrigen Impact-Faktor zitiert wird, um 62 Prozent höher ist als in einer Zeitschrift mit einem fast doppelt so hohen Impact-Faktor. Man irrt somit in 62 Prozent der Fälle, wenn man sich nach dem Impact-Faktor richtet! Der Schweizerische Nationalfonds hat jüngst die DORA-Deklaration (San Francisco Declaration on Research Assessment) unterschrieben. Danach darf die Qualität eines Aufsatzes nicht nach dem Impact-Faktor der veröffentlichenden Zeitschrift bewertet werden. Bruce Alberts,

ABB.: CHRISTIAN MURZ / FOTOCOMMUNITY; WIKIMEDIA COMMONS (2)

der Chefredaktor von „Science“, 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“. 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.

Gibt es Alternativen?

Wie kann man das Bewertungsverfahren für den wissenschaftlichen Nachwuchs verbessern und zugleich die riesigen Kosten und Zeitverzögerungen vermeiden, die das derzeitige Begutachtungsverfahren verursacht? In diesem werden die Steuerzahlerinnen und -zahler von den Zeitschriftenverlagen gleich fünffach zur Kasse gebeten: Erstens zahlt der Staat Saläre für die Verfasser der Artikel, zweitens für die Gutachter und Editoren, soweit diese ebenfalls an Universitäten beschäftigt sind. Drittens müssen heutzutage mitunter Beträge von 500 bis 1.500 US-Dollar aufgewendet werden, wenn man ein Papier einreicht. Viertens müssen die Universitätsbibliotheken Unsummen an Lizenzgebühren an ebendiese Verlage entrichten, für die die Autoren unentgeltlich schreiben, editieren und Gutachten erstellen. Schließlich müssen die Forscher, wollen sie ihr veröffentlichtes Papier online stellen, noch einmal eine Gebühr um die 1.000 US-Dollar dafür zahlen.

Der erste Vorschlag besteht darin, die Anlässe für Evaluationen auf wenige karriererelevante Entscheidungen zu reduzieren, z. B. bei der Bewerbung um eine Stelle oder bei der Beantragung von zusätzlichen Forschungsmitteln. Eine sorgfältige Eingangskontrolle ersetzt die kontinuierliche Bewertung durch dauernde Evaluationen. Sie hat die Aufgabe, das Innovationspotential, die Motivation für selbstorganisiertes Arbeiten und die Identifikation mit dem „taste of science“ zu überprüfen. Wer dieses „Eintrittsticket“ in die Gelehrtenrepublik aufgrund einer rigorosen Prüfung erworben hat, sollte weitgehende Autonomie einschließlich einer angemessenen Grundausstattung erhalten. 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 all den geschilderten Problemen angewiesen.

Hier verspricht ein offenes Post-Publication-Peer-Review-Verfahren Abhilfe. Dieses Verfahren sieht widersprüchliche Gutachten nicht als Problem, sondern als Zeichen solider und produktiver Wissenschaft. Kontroversen bieten Anlass für die Fortentwicklung der Wissenschaft, allerdings nur dann, wenn Gutachten zu einem offenen wissenschaftlichen Diskurs führen. Dies ist bei der derzeitigen Doppelt-Blind-Begutachtung nicht möglich. Im neuen Verfahren würden Forschende einen erfahrenen Kollegen oder eine Kollegin als „Editor“ anfragen, ob er oder sie Kommentare einholt, die auf einer gemeinsamen Plattform veröffentlicht werden. Die Stellungnahmen sollten mit Namen gekennzeichnet sein und können als kleine zitierfähige und reputationswirksame Veröffentlichungen gelten. Die Verfasser des ursprünglichen Artikels können auf derselben Plattform antworten. Nur wenn ein lebendiger Diskurs zustandekommt, ist

DIE AUTORIN

Prof. Dr. Margit Osterloh ist em. Professorin für Betriebswirtschaftslehre an der Universität Zürich. Ihre Spezialgebiete in Forschung und Lehre sind u. a. Organisations- und Unternehmenstheorien, Innovations- und Technologiemanagement, Knowledge Management sowie Gender Economics.

der ursprüngliche Aufsatz wissenschaftlich ergiebig. Erhält ein Papier keinen oder wenige Kommentare, signalisiert dies mangelhafte Qualität bzw. wissenschaftliche Relevanz. Sind die Kommentare oberflächlich oder gar feindselig (wie dies bei anonymen Gutachten allzu häufig der Fall ist), schädigt dies die Reputation des Gutachtenden. Vielmehr entsteht infolge der Transparenz ein 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 Begutachtungswesen endlich in das Internetzeitalter führen. Es beseitigt das Platzproblem, weil im Internet unbeschränkt viel Raum für Publikationen zur Verfügung steht. Es kann viel schneller arbeiten als das träge heutige Begutachtungssystem, bei dem mitunter zwei Jahre von der Einreichung bis zur Veröffentlichung verstreichen. Bei interessanten Papieren wäre eine rasche Rückkopplung zu erwarten. Darüber hinaus erspart es Steuerzahlerinnen und Steuerzahlern die immensen Kosten, welche ihnen die Verlage heute auferlegen. Das Verfahren lädt deutlich weniger zu einem „Gaming the System" ein. Entscheidend ist jedoch: Argumentativer Diskurs in der „Republic of Science" erhält wieder Vorrang gegenüber quantitativen Kriterien, also Zählübungen wie Impact-Faktoren und Rankings.

Die Durchsetzung dieses neuen Verfahrens wäre nicht einfach. Neben Gewinnern (dem wissenschaftlichen Nachwuchs) gibt es auch Verlierer (vor allem Verlage). Auch dürften Einrast- oder Lock-in-Effekte eintreten, die den Übergang erschweren. Aber angesichts der riesigen Probleme des heutigen Systems wäre zu wünschen, dass endlich eine Diskussion über Alternativen stattfindet.

Die drei großen Abbildungen stammen vom Fotokünstler Christian Murz.

ABB.: CHRISTIAN MURZ / FOTOCOMMUNITY

Ausgewählte Literatur

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

B. S. Frey, M. Osterloh, Schlechte Behandlung des wissenschaftlichen Nachwuchses und wie man das ändern könnte, in: Ökonomenstimme, 28. Oktober 2014.

M. Osterloh, B. S. Frey, Ranking Games und wie man sie überwinden kann, in: Zeitschrift für Kulturwissenschaft, im Druck (2015).

M. Osterloh, A. Kieser, 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, Cham et al. 2015, 307–324.

S. Ringelhan, J. Wollersheim, I. M. Welpe, Performance Management and Incentive Systems in Research Organizations: Effects, Limits and Opportunities, in: I. M. Welpe, J. Wollersheim, S. Ringelhan, M. Osterloh (Hrsg.), Incentives and Performance – Governance of Research Organizations, Cham et al. 2015, 87–106.