Psychologie

Werden die Schweine vom Wiegen fetter?

Evaluation in der Wissenschaft: Wie man sich darauf einstellt.

Von Jürgen Kaube

WENN MESSUNGEN KRITISIERT werden, wird oft das Sprichwort „Vom Wiegen werden die Schweine nicht fetter" aufgerufen. Dabei führt es in sozialen Zusammenhängen leicht in die Irre. Schon für das Schweinewiegen gilt schließlich, dass die Kreatur gemästet, weil sie gewogen und gewogen, weil sie je nach dem Ergebnis anders bepreist wird. Vom verkaufsorientierten Wiegen werden manche Schweine also durchaus fetter, als sie sonst wären.

Die Folge der Evaluation: Reaktivität

Die Forschung zu sozialen Zahlen spricht hier von „Reaktivität", um den Einfluss des Messens auf das Messergebnis zu bezeichnen. Personen verändern ihr Verhalten, je nachdem ob und wie sie geprüft und bewertet werden, vor allem dann, wenn an das Prüfergebnis Entscheidungen anschließen. Das „teaching to the test", also ein Unterricht, der sich an dem orientiert, was „klausurrelevant" ist, bietet dafür ein Beispiel.

Wenn es solche Verhaltensveränderungen gibt, steht die Einführung von Evaluationsverfahren vor der Frage, ob diese überhaupt messen, was sie messen wollen: Messen sie beispielsweise Leistungsfähigkeit oder Anpassungsfähigkeit? Beides muss, wie die Schulklausuren zeigen, einander nicht ausschließen. Herauszufinden, was von den Prüfern erwartet wird, und sich entsprechend anzupassen, ist auch eine kognitive Leistung. Doch es liegt auf der Hand, dass es nicht immer dieselbe ist. Die Vorstellung neutraler, objektiver Messung jedenfalls lässt sich nicht halten, wenn es Reaktivität gibt, und zur Frage, was überhaupt gemessen wird, tritt diejenige hinzu, ob es sich bei den durch Evaluation angeregten Verhaltensveränderungen um erwünschte oder unerwünschte handelt.

Evaluationen finden in der Wissenschaft und vor allem im Bereich der Hochschulen inzwischen auf allen Ebenen statt. Es gibt Indikatoren für die Leistung von einzelnen Forschern und von ganzen Fachbereichen, obwohl diese keine Kollektivakteure sind, die Lehre wird evaluiert, Blaupausen für Studiengänge werden es, und sogar die Qualität ganzer Universitäten findet sich in Rangordnungen gebracht. Jeder dieser Indikatoren ist hochumstritten. Bei so gut wie keinem Indikator hat wissenschaftliche Kritik daran aber bewirkt, dass er nicht mehr verwendet würde. Das könnte dafür sprechen, dass sich die Wissenschaft auf die Verwendung auch fragwürdiger Kennziffern eingestellt hat oder dass zumindest die Folgen ihrer Verwendung nicht unumgänglich und nicht eindeutig sind.

Beispiel 1: Zitationsanalyse

Betrachten wir zwei der wichtigsten Indikatoren bei Evaluationen wissenschaftlicher Leistung im deutschen Wissenschaftssystem: Zitationsanalyse und Drittmitteleinkommen. Die erste, international verbreitete Methode, misst ihrem Anspruch nach den Einfluss, den ein Wissenschaftler mit seinen Publikationen auf den Erkenntnisgewinn seiner Kollegenschaft besitzt. Die zweite, die man einen „deutschen Sonderweg" der Messung von Forschungsleistungen genannt hat (Gerhards 2013), schließt aus der Höhe gewährter Fremdfinanzierung auf die Kreditwürdigkeit wissenschaftlicher Unternehmer, interpretiert also den wiederholt gewährten Input als gutes Signal für die Qualität des Outputs. Welche Art von Reaktivität ist bei beiden Verfahren zu erwarten?

Nehmen wir zunächst bibliometrische Kriterien wie Zitationsmaße. Hier wird belohnt, wer oft und in prominenten Publikationen zitiert wird und wer kontinuierlich publiziert. Der vielverwendete Hirsch-Index beispielweise, der

ABB.: ALLKINDZA/ISTOCKPHOTO.COM

Zitationen füttern sowohl die eigene Reputation als auch die Druckindustrie.

ABB.: LUOMAN/ISTOCKPHOTO.COM; UNIVERSITY OF CHICAGO

Rund 300.000 wissenschaftliche Veröffentlichungen aus Deutschland allein 2013 – wie verändern Zitationsmaße wie der Hirsch-Index das Publikationsverhalten in Wissenschaft und Forschung?

beansprucht, die Leistungsfähigkeit eines Forschers in einer einzigen Zahl auszudrücken, hält den Zusammenhang von Zitiertwerden und Menge der Publikationen fest. Wenn von 100 Publikationen eines Wissenschaftlers 20 mindestens zwanzig Mal zitiert worden sind, die 21. auf der Rangliste des Zitiertwerdens aber weniger oft, hat der Forscher einen Hirsch-Index von 20. Wenn so gemessen wird, genügt es also nicht, ein, zwei bahnbrechende Artikel zu publizieren, um auf eine höhere Wertung als eine Person zu kommen, die ständig Beiträge verfasst, die anderen nützlich sind. Das erscheint insofern sinnvoll, als Autoren, die wenig, aber überragend publizieren, in der Frage nach ihrem Ruf ohnehin nicht auf Kennzahlen angewiesen sind. Quantitative Signale werden vielmehr gerade dann herangezogen, wenn die Forscher sich im Mittelfeld der Disziplin befinden und – in Berufungskommissionen oder bei der Mittelvergabe – andere eindeutige Vergleichskriterien fehlen.

Die Wahrscheinlichkeit, oft zitiert zu werden, hängt dabei von sehr verschiedenen Faktoren ab: von Qualität und Menge der eigenen Veröffentlichungen, von der Menge der Publikationen und der Zahl der Zeitschriften auf dem betreffenden Spezialfeld, von der Existenz prominenter Zeitschriften, denen „Impact“ zugeschrieben wird, von der Dienlichkeit des Beitrages, von der Prominenz des Autors (des Labors, der Forschungsgruppe) und von seinem Vernetzungsgrad sowie nicht zuletzt vom Verhalten von Zeitschriften-Editoren, die bekanntermaßen oft auf bestimmte Zitationen drängen.

Die Anpassungsleistungen, die eine solche Evaluation nahelegt, sind entsprechend vielfältig. So sind inzwischen gut zehn Prozent aller Zitationen Selbstzitationen.[1] Sie zahlen sich für die Autoren aus. Denn auch wenn sie nicht in die Zitationswertung des Artikels eingehen, in dem das Selbstzitat erfolgte, generieren sie mit einer gewissen Wahrscheinlichkeit Zitationen von anderen Beiträgen desselben Autors. Das liegt daran, dass Zitate in vielen Fällen nicht mehr auf Lektüre beruhen, sondern Abschriften von Bibliographien und Fußnoten sind. In unserer ersten Fußnote unten etwa haben wir nur den ersten Aufsatztitel gelesen, der selbst auf die weiteren mit einem nicht sehr verpflichtenden „vgl.“ verweist. Erkennen könnte man das hier an den nicht aufgeschlüsselten Abkürzungen der Vornamen, aber, Hand aufs Herz, wer hat das jetzt bemerkt?

1 So James H. Fowler, Dag W. Aksnes: „Does self-citation pay?“, *Scientometrics* Vol. 72 No. 3 (2007), S. 427–437 (434); vgl. Dag W. Aksnes: „A macro study of self-citation”, *Scientometrics*, 56 (2003), S. 235–246.; H. Snyder, S. Bonzi: „Patterns of self-citation across disciplines (1980–1989)”, *Journal of Information Science*, 24 (1998), S. 432–435 und R. Tagliacozzo: „Self-citation in scientific literature”, *Journal of Documentation*, 33 (1977), S. 251–265.

Mit anderen Worten: Die Zitationsanalyse erfolgt im Kontext wissenschaftlicher Praktiken, die gerade keine sehr weitgehenden Schlüsse von Zitaten auf Reputationszuweisung erlauben. Darum kann die Publikation als Litfaßsäule für eigene Beiträge, aber auch für die von Netzwerkteilnehmern oder zur rhetorischen Demonstration von Fleiß genutzt werden, und das kann erfolgreich sein, weil sich das Zitierverhalten von der Lektüre abgekoppelt hat. Der amerikanische Soziologe Andrew Abbott hat am Beispiel einer eigenen Monographie ermittelt, wie viele Zitationen auf sie verweisen, ohne dass dem ein erkennbarer Bezug auf die Argumente seines Buches zu entnehmen wäre. In zehn Prozent aller Fälle wurde es für das Gegenteil von dem zitiert, was in ihm behauptet wird. Drei Viertel aller Zitate verwiesen pauschal auf das Buch, ohne eine einzelne Passage anzugeben. Vielfach wird es dabei für Thesen zitiert, die von Autoren stammen, die Abbott seinerseits anführt.[2]

Signale, das lehrt die Informationsökonomie, sind hilfreich, wenn es teuer ist, sie zu produzieren. Evaluationen, die auf Zitatanalysen beruhen, sollten sich fragen, wie kostspielig es wirklich ist, Artikel zu verfassen, die ausreichend oft zitiert werden, und wie aufwändig es ist, Fußnoten zu füllen.

Beispiel 2: Drittmitteleinkommen

Ähnliches gilt für das Kriterium der Drittmittel. Denn welche Reaktionsmuster sind zu erwarten, wenn ihre Höhe zum zentralen Symbol von Leistung überhaupt wird? Der Soziologe Richard Münch hat darauf hingewiesen, dass Drittmittel eine Funktion von Größe sind: Je größer ein Fachbereich, desto stärker steigen die Chancen, Drittmittelprojekte einzuwerben. Das wiederum hängt mit der Arbeitsteilung zusammen, die nötig ist, um größere Projekte überhaupt antragsreif zu machen. Also differenziert das System immer mehr Rollen aus, die sich mit Meta-Aktivitäten der Forschung befassen: Sprecherrollen, Wissenschaftsmanagement, Antragsschreiben. Jeder weiß inzwischen, dass die Erfolgs-, also Finanzierungschancen eines Antrags und die Erkenntnischancen der entsprechenden Forschung zweierlei sind, weil das Antragsverfahren eine eigene Rhetorik und ein eigenes „window-dressing“ verlangt. Wenn nicht Forschung honoriert wird, sondern Anträge, verlagert sich eben ein Großteil der Arbeit auf sie. Außerdem sind die Netzwerkeffekte dieser Art von Evaluation beträchtlich, denn die Wissenschaftler treten bei der Drittmittelvergabe in der Rolle der Gutachter wie der Begutachteten auf. Mithin lohnt sich es sich, um es salopp zu formulieren, eine Betriebsnudel zu sein, denn das erzeugt Reziprozitätspflichten.

Das alles kann man wollen – Ausdifferenzierung von Managementrollen, starke Vernetzung, Größenwachstum der Projekte, erhöhte Sensibilität für das, was gerade der Trend ist, Kumulation der reputierten Forscher an wenigen Standorten, Stratifikation der Universitäten. Wenn man das will, ist Dauerevaluation ein Mittel, es zu befördern. Demgegenüber stehen ihre Kosten, in Form von zeitlichem Aufwand, in Form von Konformitätseffekten und in Form von Illusionen. Denn Reputation, die nicht auf Lektüre beruht, und Projekte, deren Abschlussberichte die Zweitfassungen ihrer Beantragung sind, können leicht etwas Illusorisches haben. Manche Schweine werden vom Wiegen schon schwer, aber in der Pfanne schnurrt ihr Fleisch dann mitunter erstaunlich schnell zusammen.

Der amerikanische Soziologe Andrew Abbott hat am Beispiel seiner eigenen Monographie ermittelt, wie viele Zitationen auf sie verweisen, ohne dass dem ein erkennbarer Bezug auf die Argumente seines Buches zu entnehmen wäre.

DER AUTOR

Jürgen Kaube ist Ressortleiter für die „Geisteswissenschaften“ der Frankfurter Allgemeinen Zeitung.

2 Andrew Abbott: „Varieties of Ignorance”, The American Sociologist, Vol. 41, 2 (2010), S. 174–189.