Bibliometrie: Das Leidener Manifest zu Forschungsmetriken

Diana Hicks, Paul Wouters, Ludo Waltman, Sarah de Rijcke, Ismael Rafols

Nature 520, 429–431 (23. April 2015), doi:10.1038/520429a

Übersetzt von Gerald Langhanke, Universitäts- und Landesbibliothek Darmstadt, orcid.org/0000-0001-9964-1112 (13. August 2015)

Nutzen Sie diese zehn Grundsätze um Forschung zu bewerten, drängen Diana Hicks, Paul Wouters und Kollegen.

Daten werden immer stärker zur Steuerung von Forschung genutzt. Forschungsevaluationen, die einst von Fachkollegen bezeugt und durchgeführt wurden, sind heute alltäglich und abhängig von Metriken[i]. Das Problem ist, dass diese Evaluationen heute von den Daten bestimmt werden, anstelle des Urteilsvermögens. Die Metriken sind gewuchert: meistens gut gemeint, nicht immer fundiert, häufig falsch angewendet. Wir riskieren das System mit den Werkzeugen zu beschädigen, die dazu gedacht waren es zu verbessern, da Forschungsbewertungen immer öfter von Organisationen vorgenommen werden, die über kein Wissen und keine Beratung zu anerkannten Methoden und Deutungen verfügen.

Vor 2000 gab es den *Science Citation Index* des *Institute for Scientific Information* (ISI) auf CD-ROM, der von Experten für spezielle Analysen genutzt wurde. 2002 machte *Thomson Reuters* die *Web of Science*-Datenbank im Internet breit zugänglich. Konkurrierende Zitationsindizes wurden geschaffen: *Scopus* von *Elsevier* (ab 2004) und *Google Scholar* (als Beta-Version ab 2004). Web-basierte Werkzeuge zum leichten Vergleich von Produktivität und Außenwirkung institutioneller Forschung wurden eingeführt, zum Beispiel *InCites* (beruhend auf *Web of Science*) und *SciVal* (beruhend auf *Scopus*), ebenso Software, die erlaubt individuelle Zitationsprofile anhand von *Google Scholar* zu analysieren (*Publish or Perish*, ab 2007).

2005 schlug Jorge Hirsch, ein Physiker an der University of California, San Diego, den *h*-Index vor, und machte so das Zählen von Zitaten für einzelne Wissenschaftler populär. Die Bedeutung des *journal impact factor* wuchs seit 1995 beständig (siehe Grafik „Impact-factor obsession“).

In letzter Zeit nahmen Metriken Fahrt auf, die mit der sozialen Nutzung und Online-Kommentaren verbunden sind – *F1000Prime* startete 2002, *Mendeley* 2008 und *Altmetric.com* (getragen von *Macmillan Science and Education*, Eigentümer der *Nature Publishing Group*) 2011.

Als Szientometriker, Sozialwissenschaftler und Forschungsmanager haben wir mit wachsender Beunruhigung den allgegenwärtigen Missbrauch der Indikatoren zur Evaluation wissenschaftlicher Leistungen beobachtet. Die folgenden sind nur wenige

Beispiele von vielen. Überall auf der Welt sind Universitäten fixiert auf ihr Abschneiden in globalen Rankings (wie zum Beispiel das *Shanghai Ranking* und die *Times Higher World University Rankings*), auch wenn diese Listen auf, in unseren Augen, ungenauen Daten und beliebigen Indikatoren beruhen.

Manche Personalabteilungen verlangen *h*-Index-Werte für Bewerber. Einige Universitäten stützen Promotionsentscheidungen auf Schwellwerte für den *h*-Index und auf die Anzahl der Artikel in *high-impact*-Zeitschriften.

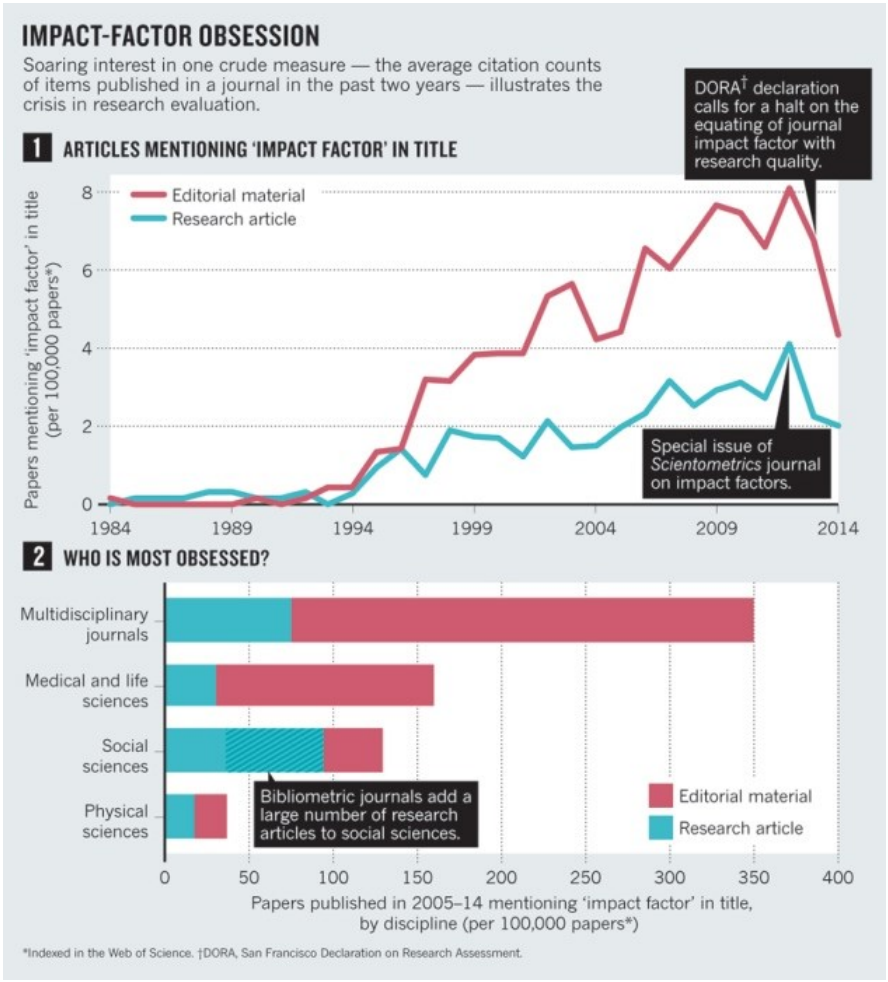
Die Lebensläufe von Wissenschaftlern sind Gelegenheiten geworden mit diesen Werten zu prahlen, besonders in der Biomedizin. Überall fordern Doktorväter und -mütter ihre Doktoranden auf in *high-impact*-Zeitschriften zu veröffentlichen und Drittmittel einzuwerben bevor diese dazu bereit sind.

In Skandinavien und China vergeben manche Universitäten Forschungsmittel und Boni aufgrund einer Zahl: beispielsweise durch die Berechnung eines individuellen *impact factors* um „leistungsbezogene Mittel" zu verteilen oder durch eine Bonusvergabe an Forscher für eine Veröffentlichung in einer Zeitschrift mit einem *impact factor* größer als 15[ii].

In vielen Fällen legen Wissenschaftler und Gutachter ausgewogenes Urteilsvermögen an den Tag. Aber der Missbrauch von Forschungsmetriken ist zu weit verbreitet als dass er ignoriert werden dürfte.

Wir präsentieren daher das Leidener Manifest (*Leiden Manifesto*), benannt nach der Konferenz, auf der es entstand (siehe http://sti2014.cwts.nl). Seine zehn Grundsätze sind für Szientometriker nichts neues, wenngleich niemand von uns sie in Gänze aufsagen könnte, da ihre Kodifizierung bisher gefehlt hat. Von Koryphäen des Gebiets wie Eugene Garfield (Gründer des ISI) ist bekannt, dass sie einige dieser Grundsätze teilten[iiiiv]. Aber sie sind nicht anwesend, wenn Gutachter Bericht an die Mitarbeiter in der Universitätsverwaltung erstatten, die keine Experten in der relevanten Methodik sind. Wissenschaftler auf der Suche nach Literatur um ein Gutachten anzufechten, finden das notwendige Material verstreut in für sie unbedeutenden Zeitschriften, auf die sie keinen Zugriff haben.

Wir bieten dieses Destillat an vorbildlichen Praktiken für auf Metriken gestützte Forschungsbewertungen, damit Forscher ihre Gutachter in die Verantwortung nehmen können, und Gutachter ihre Indikatoren.



Datenquelle: Thomson Reuters Web of Science; Analyse: D.H., L.W.

Zehn Grundsätze

1) **Quantitative Untersuchungen sollen qualitative Bewertungen durch Experten unterstützen.** Quantitative Metriken können einseitige Tendenzen von *peer review* hinterfragen und weitere Überlegungen fördern. Dies wird den *peer review*-Prozess stärken, denn ohne eine Vielzahl an relevanten Informationen ist es schwierig Kollegen zu beurteilen. Gutachter dürfen aber nicht versucht sein ihre Entscheidung den Zahlen zu überlassen. Indikatoren sind kein Ersatz für fundierte Beurteilung. Die Verantwortung für Gutachten verbleibt bei den Gutachtern.

2) **Bewerten Sie die Leistung anhand des Forschungsziels der Institution, der Gruppe oder des Forschers.** Programmziele sollten zu Beginn festgelegt werden und die zur Leistungsmessung verwendeten Indikatoren sollten sich klar auf diese Ziele beziehen. Die Wahl von Indikatoren und die Art und Weise, in der sie verwendet werden, sollten breitere sozio-ökonomische und kulturelle Kontexte berücksichtigen. Wissenschaftler haben unterschiedliche Forschungsziele.

Grundlagenforschung an der Grenze des Wissens unterscheidet sich von Forschung, die darauf abzielt Lösungen für Fragestellungen der Gesellschaft zu entwickeln. Eine Bewertung kann auch auf Verdiensten in Bezug zur Politik, zur Wirtschaft oder zur Öffentlichkeit beruhen, nicht nur auf akademischen Exzellenzvorstellungen. Kein einzelnes Bewertungsmodell passt für alle Kontexte.

3) **Schützen Sie die Spitzenleistungen der ortsbezogenen Forschung**. In vielen Teilen der Welt wird wissenschaftliche Spitzenleistung mit englischsprachiger Veröffentlichung gleichgesetzt. Die spanische Rechtswissenschaft beispielsweise erklärt es für wünschenswert, dass spanische Wissenschaftler in *high-impact*-Zeitschriften veröffentlichen. Der *impact factor* wird für Zeitschriften berechnet, die im US-bezogenen und immer noch größtenteils englischsprachigen *Web of Science* ausgewertet werden. Diese Tendenzen sind besonders problematisch in den Sozial- und Geisteswissenschaften, in denen die Forschung stärker regional und national ausgerichtet ist. Viele weitere Disziplinen haben eine nationale oder regionale Dimension – zum Beispiel HIV-Epidemiologie im subsaharischen Afrika.

   Diese Vielfalt und gesellschaftliche Relevanz tendiert dazu verdrängt zu werden um Aufsätze zu produzieren, die für die Torwächter des *high-impact* von Interesse sind: englischsprachige Zeitschriften. Diejenigen spanischen Soziologen, die in *Web of Science* häufig zitiert sind, haben mit abstrakten Modellen oder US-amerikanischen Daten gearbeitet. Die Ausprägungen der Soziologie in einflussreichen spanischsprachigen Aufsätzen geht so verloren: Themen wie lokales Arbeitsrecht, familiäre Altenpflege oder die Arbeitslage von Immigranten[v]. Metriken, die auf hochwertiger, nicht-englischsprachiger Literatur aufbauen, würden der Identifikation und Belohnung von ortsbezogen relevanter Forschung dienen.

4) **Gestalten Sie die Sammlung von Daten und die Verarbeitungsschritte offen, transparent und einfach.** Der Aufbau der Datenbanken, die für die Evaluation benötigt werden, sollte klaren Regeln folgen, die vor Beginn der Untersuchungen festgelegt wurden. Dies war das übliche Vorgehen in den wissenschaftlichen und kommerziellen Gruppen, die bibliometrische Untersuchungsmethoden über viele Jahrzehnte entwickelt haben. Diese Gruppen bezogen sich auf Vorschriften, die in qualitätsgesicherter Literatur festgehalten wurde. Diese Transparenz ermöglichte genaue Überprüfungen. 2010 zum Beispiel führte eine öffentliche Diskussion über die technischen Eigenschaften eines wichtigen Indikators, der von einer unserer Gruppen (dem *Centre for Science and Technology Studies* an der Universität Leiden in den Niederlanden) verwendet wurde, zu einer Überarbeitung in der Berechnung dieses Indikators.[vi] Neue kommerzielle Anbieter sollten an den gleichen Standards gemessen werden, niemand sollte Blackbox-Programme für Evaluationen akzeptieren.

   Die Einfachheit eines Indikators ist eine Tugend, da sie die Transparenz fördert. Aber zu einfache Metriken können das Ergebnis verfälschen (siehe Grundsatz 7).

Gutachter müssen ein Gleichgewicht anstreben – einfache Indikatoren, die der Komplexität des Forschungsprozesses gerecht werden.

5) **Erlauben Sie den Begutachteten die Daten und deren Analyse nachzuprüfen.** Um Datenqualität zu gewährleisten, sollte es allen Forschern, die in bibliometrische Analysen einbezogen werden, ermöglicht werden zu prüfen, ob ihre Veröffentlichungen korrekt erkannt wurden. Jeder, der Evaluationsprozesse leitet und steuert, sollte die Fehlerfreiheit der Daten durch eigene Prüfung oder Prüfung durch Dritte sicherstellen. Universitäten können dies in ihren Forschungsinformationssystemen umsetzen; dies sollte eine Leitlinie bei der Auswahl von Anbietern dieser Systeme sein. Korrekte und qualitativ hochwertige Daten zu sammeln und zu verarbeiten kostet Zeit und Geld. Dies muss im Haushalt vorgesehen werden.

6) **Berücksichtigen Sie die unterschiedlichen Publikations- und Zitierungskulturen.** Es ist eine bewährte Methode eine Menge von möglichen Indikatoren auszuwählen und den Fachdisziplinen zur Auswahl zu stellen. Vor wenigen Jahren erhielt eine Gruppe europäischer Historiker eine vergleichsweise schlechte Beurteilung in einer nationalen *peer review*-Begutachtung, da sie eher Bücher statt Aufsätze in von *Web of Science* ausgewerteten Zeitschriften veröffentlichte. Den Historikern wurde zum Verhängnis, dass sie zum Fachbereich Psychologie gehörten. Bei Geistes- und Sozialwissenschaftlern ist es notwendig Bücher und landessprachliche Literatur in die Zählung von Veröffentlichungen aufzunehmen; bei Informatikern müssen Konferenzbeiträge berücksichtigt werden.

   Zitationsraten variieren je nach Disziplin. Die höchst-platzierten Zeitschriften in der Mathematik haben *impact factors* von etwa 3, diejenigen in der Zellbiologie solche von an die 30. Es werden normierte Indikatoren benötigt und die zuverlässigste Normierungsmethode beruht auf Perzentilen: Jeder Aufsatz wird mit dem Perzentil gewichtet, zu dem es in der Zitationsverteilung seiner Disziplin gehört (z.B. die Top-1%, 10% oder 20%). Eine einzelne, vielzitierte Veröffentlichung verbessert die Position einer Universität nur geringfügig in einer Rangfolge, die auf Perzentil-Indikatoren beruht, aber treibt die Universität aus dem Mittelfeld in die Spitzengruppe in einer Rangliste, die auf Durchschnitts-Zitationsraten beruht.[vii]

7) **Gründen Sie die Beurteilung von einzelnen Forschern auf eine qualitative Einschätzung ihrer Veröffentlichungsliste.** Je älter man ist, desto höher ist der *h*-Index, auch bei ausbleibenden neuen Aufsätzen. Der *h*-Index variiert von Feld zu Feld: Lebenswissenschaftler erreichen bis zu 200, Physiker an die 100 und Sozialwissenschaftler 20-30[viii]. Diese Zahlen sind abhängig von der verwendeten Datenbank: es gibt Forscher in Informatik, die einen *h*-Index von ca. 10 bei *Web of Science* haben, aber 20-30 bei *Google Scholar*[ix]. Die Arbeiten eines Forschers zu lesen und einzuschätzen ist sehr viel angemessener als sich auf eine einzige Zahl zu verlassen. Auch im Vergleich großer Anzahlen von Forschern ist ein

Ansatz, der mehr Informationen über die persönliche Expertise, Erfahrung, Aktivitäten und Einflüsse berücksichtigt, der beste.

8) **Vermeiden Sie unpassende Konkretheit und falsche Genauigkeit.** Indikatoren für Wissenschaft und Technik sind anfällig für konzeptionelle Mehrdeutigkeiten und Unsicherheiten und setzen starke Annahmen voraus, die nicht allgemein akzeptiert sind. Die Bedeutung von Zitatzählungen zum Beispiel wurde lange diskutiert. Daher sollten am besten mehrere Indikatoren verwendet werden um einen zuverlässigeren und vielfältigeren Eindruck zu erhalten. Wenn Unsicherheiten und Fehler quantifiziert werden können (z. B. durch Fehlerbalken), sollten diese Informationen den veröffentlichten Werten beigefügt werden. Falls dies nicht möglich ist, sollten die Ersteller zumindest falsche Genauigkeit vermeiden. Der *journal impact factor* zum Beispiel wird mit drei Nachkommastellen veröffentlicht, um Gleichstände zu vermeiden. Betrachtet man aber die konzeptionelle Mehrdeutigkeit und die zufälligen Schwankungen von Zitatzählungen, ergibt es keinen Sinn zwischen Zeitschriften aufgrund sehr kleiner Unterschiede im *impact factor* zu unterscheiden. Vermeiden Sie falsche Genauigkeit: nur eine Nachkommastelle ist gerechtfertigt.

9) **Erkennen Sie die systematischen Effekte der Gutachten und Indikatoren.** Indikatoren verändern durch die von ihnen gesetzten Anreize das System. Diese Auswirkungen sollten verhindert werden. Daher ist eine Sammlung von Indikatoren immer vorzuziehen – ein einzelner Indikator lädt zu Spielerei und zu Zielverlagerungen ein (sodass das die Messung zum Ziel wird). In den 1990ern förderte zum Beispiel Australien die universitäre Forschung anhand einer Formel, die im Wesentlichen auf der Anzahl der von einer Institution veröffentlichten Aufsätze beruhte. Universitäten konnten den „Wert“ eines in einer begutachteten Zeitschrift veröffentlichten Aufsatzes ausrechnen; im Jahr 2000 betrug er 800 AU$ (2000 etwa 500 €) an Forschungsförderung. Wie vorherzusehen war, nahm die Zahl der von australischen Forschern veröffentlichten Aufsätze zu. Dies aber in wenig-zitierten Zeitschriften, was auf fallende Qualität der Artikel hindeutet[x].

10) **Hinterfragen und aktualisieren Sie die Indikatoren regelmäßig.** Forschungsfragen und die Ziele ihrer Begutachtung ändern sich und das Wissenschaftssystem selbst entwickelt sich mit ihnen. Einst nützliche Metriken werden unpassend und neue entstehen. Indikatorensammlungen müssen überprüft und gegebenenfalls angepasst werden. Als die Effekte einer grob vereinfachenden Formel erkannt wurden, führte Australien 2010 die komplexere *Excellence in Research for Australia initiative* durch, die mehr Wert auf qualitative Untersuchungen legt.

Die nächsten Schritte

Durch Einhaltung dieser zehn Grundsätze kann Forschungsevaluation eine wichtige Rolle in der Weiterentwicklung von Wissenschaft und ihrer Wechselwirkung mit der Gesellschaft spielen. Forschungsmetriken können bedeutende Informationen liefern, die durch persönliche Expertise nur schwierig zu gewinnen oder zu verstehen sind. Diese quantitativen Informationen dürfen sich nicht von einem Werkzeug in das Ziel verwandeln.

Die besten Entscheidungen werden durch die Verbindung von zuverlässiger und zweckmäßiger Statistik mit der jeweiligen Forschungskultur getroffen. Sowohl quantitative wie qualitative Belege werden benötigt; beide sind auf ihre Art objektiv. Entscheidungen in der Wissenschaft müssen sich auf gesicherte Prozesse stützen, die auf höchstmöglicher Datenqualität beruhen.

---

[i] Wouters, Paul, in *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*, hg. von Blaise Cronin und Cassidy R Sugimoto (MIT Press, 2014), 47–66.

[ii] Jufang Shao und Huiyun Shen, „The Outflow of Academic Papers from China: Why Is It Happening and Can It Be Stemmed?“, *Learned Publishing* 24, Nr. 2 (1. April 2011): 95–97, doi:10.1087/20110203.

[iii] P. O Seglen, „Why the Impact Factor of Journals Should Not Be Used for Evaluating Research“, *BMJ* 314, Nr. 7079 (15. Februar 1997): 497–497, doi:10.1136/bmj.314.7079.497.

[iv] Ebd.; Garfield E, „The history and meaning of the journal impact factor“, *JAMA* 295, Nr. 1 (4. Januar 2006): 90–93, doi:10.1001/jama.295.1.90.

[v] C. Lopez Pineiro und D. Hicks, „Reception of Spanish Sociology by Domestic and Foreign Audiences Differs and Has Consequences for Evaluation“, *Research Evaluation* 24, Nr. 1 (1. Januar 2015): 78–89, doi:10.1093/reseval/rvu030.

[vi] Anthony F.J. van Raan u. a., „Rivals for the Crown: Reply to Opthof and Leydesdorff“, *Journal of Informetrics* 4, Nr. 3 (Juli 2010): 431–35, doi:10.1016/j.joi.2010.03.008.

[vii] Ludo Waltman u. a., „The Leiden Ranking 2011/2012: Data Collection, Indicators, and Interpretation“, *Journal of the American Society for Information Science and Technology* 63, Nr. 12 (Dezember 2012): 2419–32, doi:10.1002/asi.22708.

[viii] J. E. Hirsch, „An Index to Quantify an Individual's Scientific Research Output“, *Proceedings of the National Academy of Sciences* 102, Nr. 46 (15. November 2005): 16569–72, doi:10.1073/pnas.0507655102.

[ix] Judit Bar-Ilan, „Which H-Index? — A Comparison of WoS, Scopus and Google Scholar“, *Scientometrics* 74, Nr. 2 (Februar 2008): -, doi:10.1007/s11192-008-0216-y.

[x] Linda Butler, „Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts“, *Research Policy* 32, Nr. 1 (Januar 2003): 143–55, doi:10.1016/S0048-7333(02)00007-0.