BUNSEN-MAGAZIN · 20. JAHRGANG · 6/2018

Martin Quack

ZUR GUTEN PRAXIS DER EVALUATION WISSENSCHAFTLICHER FORSCHUNG

Die internationale Entwicklung in der wissenschaftlichen Forschung hat zu einem außerordentlichen Wachstum sowohl in der Zahl der Forscherinnen und Forscher als auch in den finanziellen Mitteln geführt, die von den Staaten hierfür zur Verfügung gestellt werden. In diesem Zusammenhang gibt es auch zunehmende Forderungen nach Rechtfertigung der Mittelverwendung durch geeignete Evaluationen der geförderten Forschung. Dabei geht es insbesondere auch um die Bewertung der Leistungen von Forscherinnen und Forschern. Drei bedeutende nationale Akademien der Wissenschaften in Europa, die Académie des Sciences in Frankreich, die Leopoldina in Deutschland und die Royal Society in Großbritannien haben gemeinsam aus Sorge um Fehlentwicklungen bei Evaluationsverfahren und insbesondere wegen der missbräuchlichen Verwendung von quantitativen Indikatoren wie Impakt Faktoren (IF, impact factor) Leitlinien für eine verantwortungsvolle, gute Evaluationspraxis ausgearbeitet. Diese wurden in einer ausführlichen Diskussion im Dezember 2017 dem Kommissar für Forschung, Wissenschaft und Innovation bei der Europäischen Union Carlos Moedas in Brüssel vorgestellt, der das Dokument zustimmend aufgenommen hat. Wegen der Bedeutung des Gegenstandes drucken wir im Anhang (S. 209-212) das bewusst kurz gehaltene Dokument im englischen Wortlaut ab und ergänzen es hier nach einer kurzen Zusammenfassung der Hauptpunkte mit einigen Gedanken zur Bewertung wissenschaftlicher Forschung, wie sie bei der Förderung von Forschungsprojekten oder von Forscherinnen und Forschern und auch etwa bei Berufungen und Beförderungen notwendig ist. Eine Bewertung – mit allerdings weiteren Besonderheiten – wird auch bei der Vergabe von Preisen vorgenommen und wir weisen auf einen früheren Leitartikel zu diesem Thema hin, da auch in diesem Jahr wieder einmal die Termine für Preisnominationen bei der Bunsen-Gesellschaft anstehen [1]. Auch auf die früher von den drei Akademien publizierten Leitsätze für eine gute wissenschaftliche Publikationspraxis sei hier verwiesen, da sie ergänzende Aussagen zur Evaluation wissenschaftlicher Publikationen enthalten [2].

Prof. Dr. Martin Quack
Laboratorium für Physikalische Chemie
ETH Zürich
CH-8093 Zürich, Schweiz
E-Mail: martin@quack.ch

* Richard R. Ernst zum 85. Geburtstag gewidmet

Grundsatz jeder Bewertung soll eine inhaltliche Begutachtung der Qualität durch ausgewiesene Experten höchster Integrität sein. Bibliometrische Indices und ähnliche quantitative Indikatoren sind kein Ersatz für eine inhaltliche Bewertung, ihre häufige missbräuchliche Verwendung ist besorgniserregend und sollte als „schlechte Praxis“ (bad practice) gebrandmarkt werden. Insbesondere die Verwendung von Impakt Faktoren (IF) von Zeitschriften ist problematisch, wie auch schon in der San Francisco Declaration [3] festgehalten wurde, die von zahlreichen Institutionen unterstützt wird. Diese rein quantitativen Indikatoren enthalten keinerlei Qualitätsaussage, führen zu falschen Anreizen und einer Korrumpierung des Wissenschaftssystems. Auch die neuen Indikatoren aus den sogenannten „Altmetrics“ haben völlig vergleichbare Schwächen wie Ergebnisse aus Bibliometrie und den zitationsbasierten Indices.

Aus diesen Feststellungen folgen einige generelle Prinzipien für eine gute Evaluationspraxis.

1. Der Auswahl der besten Experten für eine Evaluation kommt besondere Bedeutung zu. Sie müssen fachlich und charakterlich den höchsten Anforderungen genügen. Die Auswahl der Experten muss transparent sein.
2. Interessenkonflikte von Experten müssen offengelegt werden und absolute Vertraulichkeit muss gewahrt bleiben.
3. Es müssen Verfahren verwendet werden, die sicherstellen, dass befangene oder sonst unangemessene Gutachten von der Berücksichtigung ausgeschlossen werden.
4. Die Prozesse der Evaluation sollten gut definiert und transparent sein, mit einer genügenden Vielfalt von Experten und auch ausreichendem Wechsel zwischen Experten um Einseitigkeiten zu vermeiden.
5. Die Zahl der Evaluationen sollten auf das wirklich notwendige Maß reduziert werden, etwa auf wichtige Personalentscheidungen, Berufungen oder Entscheidungen über Förderungen mit großen Finanzvolumina.
6. Bei der Begutachtung sollen die Experten die wissenschaftliche Qualität in geeigneter Weise aus dem Studium von ausgewählten Publikationen, Büchern, Vorträgen und anderen wissenschaftlichen Leistungen erschließen. Die Vorschläge der zu Evaluierenden sollen hierzu bei der Auswahl der zu bewertenden Leistungen erfragt und berücksichtigt werden. Bei Patenten muss zwischen den verschiedenen Phasen unterschieden werden (Einreichung, Bewilligung, Nutzung). Die rein quantitativen Indikatoren wie Publikationszahl, Zahl der Zitate oder andere bibliometrische Indices und der Umfang eingeworbener Forschungsmittel sollten untergeordnete Bedeutung bei der Bewertung haben. *Die wesentlichen Kriterien sind Qualität, Originalität und Bedeutung der Forschung.**

Diese kurze Zusammenfassung wesentlicher Punkte aus dem Dokument der Akademien soll hier noch mit einigen Kommentaren ergänzt werden. Ein häufiger Einwand gegen die von den Akademien stark hervorgehobene und befürwortete Begutachtung der wissenschaftlichen Qualität durch anerkannte Experten ist die befürchtete Einseitigkeit („old boys clubs" oder „Schulen"). Dem wird durch die geforderte Diversifizierung, Vielseitigkeit und heterogene Zusammensetzung von Begutachtungsgremien und ausgewählten Gutachtern begegnet, alles auf hohem internationalen Niveau. Hierdurch werden unangemessene Begutachtungen Ausnahmen, die von Fall zu Fall identifiziert und in der Bewertung ausgeschlossen werden. Nach meiner Erfahrung als Mitglied entsprechender nationaler und internationaler Gremien und auch als Vorsitzender von Berufungskommissionen (als Delegierter des Präsidenten für Wahlkommissionen, DPW während 17 Jahren für viele Departemente der ETH, aus Prinzip nie im eigenen Departement) kommen die erwähnten negativen Erscheinungen mit Falschbegutachtung und schulenmäßiger Einseitigkeit durchaus gelegentlich vor, wenn auch seltener als oft vermutet. Sie können aber durch geeignete Verfahren korrigiert werden. Es gehört zum Ethos der Begutachtung, dass Befangenheit ebenso spontan offengelegt wird wie auch mangelnde Kompetenz bei einem speziellen Forschungsbereich, mit entsprechendem Verzicht auf Begutachtung in solchen Fällen. Gegen diese klare Regel wird nicht sehr oft verstoßen.

Ein häufigeres Fehlverhalten von Experten beruht eher auf Nachlässigkeit (wegen Trägheit oder „allgemeiner Überlastung"). Gerade dann greifen die „Experten" auf Argumente wie Zahl der Publikationen in HIF-Zeitschriften (High-Impact Factor) oder bibliometrische Indices zurück. Nach solchen Argumenten stellte ich als definitionsgemäß neutraler Vorsitzender und Nichtexperte die Frage an die betreffenden Experten: „Bitte erläutern Sie doch der Kommission und auch für mich als Fachfremden verständlich die wissenschaftliche Bedeutung der betreffenden Arbeiten". In keinem einzigen Fall in vielen Beispielen in 17 Jahren konnte der jeweilige Experte die Frage angemessen beantworten, weil er sich eben inhaltlich gar nicht informiert hatte. Es gab dann aber stets andere Mitglieder der Kommission, die eine Diskussion in inhaltlicher Richtung dazu führen konnten. Bei der missbräuchlichen Überbewertung der Zahl von Publikationen in HIF-Zeitschriften muss man auch bedenken, dass diese bei vielen ein durchaus zweifelhaftes Ansehen genießen (ohne die Kolleginnen und Kollegen hier mit Namen zu benennen, sehr bedeutende darunter, kann ich hier Begriffe wie „wissenschaftliche Boulevardzeitschriften" oder gar „pornographic journals" zitieren). Ein gedrucktes (eher maßvolles) Zitat von G. Bodenhausen sei dennoch wörtlich wiedergegeben: „I was asked by *Science* to review a paper. In a brief spell of arrogance I wrote: The authors would be well advised to submit their work to a more serious journal less inclined to hype"[4]. Die Gefahr der Korrumpierung der Wissenschaft durch die Überbewertung von HIF-Zeitschriften, die selbst sehr stark durch kommerzielle Interessen getrieben sind, ist in neuester Zeit bekannt geworden, durch die Analyse von finanziellen „Belohnungen", die von chinesischen Universitäten an ihre Wissenschaftler für jede Publikation in einer HIF-Zeitschrift gezahlt werden (Beträge von 25'000-50'000 US-Dollar werden für einen Artikel in Science oder Nature berichtet, wobei hinzugefügt wird: „this has an impact on the behavior of some scientists, plagiarism, academic dishonesty, ghostwritten papers and fake peer review scandals are on the increase in China as is the number of mistakes)"[7]. Hierbei wird nicht eine herausragende wissenschaftliche Leistung honoriert, wie es bei einem Wissenschaftspreis der Fall ist, sondern eine „Ersatzleistung", die Publikation in einer HIF-Zeitschrift. Der Verdacht liegt nahe, dass die Universitäten hierbei das ebenfalls korrupte Ziel haben, durch den betreffenden falschen Anreiz ihre HIF Statistik zu verbessern, die ja bekanntlich für die „Rankings" der Universitäten mitbenutzt wird.

Gelegentlich wird vorgeschlagen, etwa bei der Bewertung einer sehr großen Zahl von Personen ein zweistufiges Verfahren zu verwenden, da die Gesamtzahl der zu Beurteilenden „zu groß" sei. In der Tat kann man etwa bei Berufungskommissionen durchaus 100 bis 200 oder auch größere Bewerbungszahlen erleben. Dann wird behauptet, man könne ja in einer ersten Stufe mit Hilfe schematischer Verfahren (quantitative Indikatoren etc.) die Zahl stark reduzieren, durch eine entsprechende Ausschlussgrenze bei den Indikatoren und erst in einer zweiten Stufe die nötige inhaltliche Bewertung vornehmen. Ich halte das für eine schlechte Idee. Denn dann würden in der ersten Stufe unter Umständen herausragende junge Talente gleich zu Beginn aus der Endauswahl entfernt, weil sie quasi als Aschenputtel die „quantitativen Indikatoren" nicht erfüllen. Besser ist es, in der ersten Stufe auf das Vorwissen der Experten in der Kommission zurückzugreifen, das meistens durch Kenntnis der Literatur auf dem Gebiet, herausragende Vorträge auf Tagungen etc. vorhanden ist. Ich habe als Vorsitzender jede Person, für die sich auch nur ein einziges Kommissionsmitglied bei einer Vordiskussion positiv äußerte, dann einer Detaildiskussion in der zweiten Stufe zugeführt. Weiterhin kann man die erste inhaltliche Bewertung in der ersten Stufe auf die Kommissionsmitglieder aufteilen, so dass jedes Kommissionsmitglied nur eine begrenzte Zahl von Personen zu bewerten hat, worüber dann der Kommission von dem betreffenden Mitglied berichtet wird. Nur die nach diesen Berichten inhaltlich positiv bewerteten Personen werden dann in weiteren Stufen von der gesamten Kommission im Detail diskutiert. Der Gesamtaufwand war immer im Bereich des Machbaren, ohne dass man schematische Verfahren ohne jede inhaltliche Bewertung verwenden musste.

Zwei historische Beispiele (alt und neu) können als Illustration dienen. Von einem sehr viel älteren Kollegen habe ich den Bericht von einer Berufung eines (damals) jungen Professors vor vielen Jahren an der ETH erhalten. Es wurden Gutachten eingeholt, neben positiven gab es auch ein negatives. Die Kommission schloss damals nach sorgfältiger Analyse das negative Gutachten von der Bewertung aus, die Berufung erfolgte und führte Jahrzehnte danach zu einem Nobelpreis. Zum Zeitpunkt der Verleihung des Nobelpreises wurde übrigens versucht, die alten Berufungsakten aufzufinden. Sie waren aber in den Unterlagen der ETH verlorengegangen, so dass der mündliche Bericht des alten Kollegen das einzige verbleibende Dokument ist. Ein neues Beispiel zeigt, dass eine sorgfältige inhaltliche Bewertung Berufungen herausragender Kandidaten in sehr jungem Alter ermöglicht. Die diesjährigen Fieldsmedaillengewinner Alessio Figalli (ETH) und Peter Scholze (Bonn) wurden so sehr jung (mit 25 Jahren) auf eine Professur berufen und erhielten

wenige Jahre später die Fields Medaille. Allerdings ist die akademische Tradition in der Mathematik allgemein kaum vom ‚Zitat-Indikatorvirus' infiziert, inhaltliche Beurteilung ist hier die Regel.

Von Wissenschaftsbürokraten hört man manchmal zugunsten der Zitatstatistiken: „Wenn zum Beispiel eine Arbeit über irgendeinen physikalischen Detaileffekt über Jahre hinweg kaum gelesen geschweige denn zitiert wird, dann hat sie wohl doch keinerlei Bedeutung". Es gibt unzählige Beispiele, die dies widerlegen. Hier seien nur die beiden von Einstein vorhergesagten Effekte der Gravitationslinsen und der Gravitationswellen erwähnt [8, 9]. Die betreffenden Arbeiten wurden über viele Jahrzehnte hin kaum zitiert. Heute gibt es bemerkenswerte astronomische Beobachtungen, die auf Gravitationslinsen beruhen. Auch die kürzliche Beobachtung von Gravitationswellen eröffnet ein völlig neues Fenster für die Astronomie, hat das Potential, die Grundlagen der Astronomie zu verändern und wurde kürzlich mit dem Nobelpreis ausgezeichnet. Die Bedeutung bleibt auch dann erhalten, wenn man gemäß kritischen Stimmen diesen Nobelpreis als verfrüht betrachtet wegen noch ungenügender Bestätigung der Beobachtungen. Die mangelnden Zitate über Jahrzehnte hinweg haben demgegenüber keinerlei Gewicht. Weitere historisch gut dokumentierte Beispiele sind die Experimente von Faraday zum Elektromagnetismus und die Quantenmechanik mit der Schrödingergleichung, die jeweils zu Beginn als merkwürdige physikalische Effekte oder hochspezialisierte gar spekulative theoretische Untersuchungen ohne jede praktische Konsequenz betrachtet wurden, heute aber die Grundlage großer Wirtschaftszweige in unserem täglichen Leben bilden, oft zitiert und ebenso oft wieder vergessen (siehe [10, 11]).

Was sind die qualitativen Kriterien, denen wir zusammenfassend bei der Bewertung von Forscherpersönlichkeiten und ihrer Wissenschaft Gewicht geben sollen?
1. Tiefes Erkenntnisstreben und Entdeckergeist
2. Wissenschaftliche Kreativität und Originalität auch bei der Auswahl bedeutender Forschungsprobleme
3. Wissenschaftlicher Mut auch für risikoreiche und langfristige Projekte
4. Kritischer Verstand – auch die Fähigkeit zur Selbstkritik
5. Das Feuer der Begeisterung für die Wissenschaft und eine entsprechende Kommunikationsfähigkeit besonders auch in der Lehre
6. „Guter Bürgersinn", Kollegialität, persönliche Integrität und Verantwortungsbewusstsein in der Wissenschaft [12] und darüber hinaus [13], sehr altmodisch formuliert auch Menschlichkeit und ein „gutes Herz".

Ziel der Evaluation wissenschaftlicher Forschung ist die Förderung guter Wissenschaft, nicht die Heranzüchtung von Wissenschaftsapparatschiks, die ihre „Indikatoren" optimieren.

Richard Zare hat bei seiner Auswahl der Kriterien für Berufungen an der Stanford Universität den Punkt „good departmental citizen" sogar an die erste Stelle gesetzt (neben der Qualität in Lehre und Forschung) [12], was man durch die Konsequenzen von „bad citizenship" verstehen kann (siehe die Diskussion in [11]). Ziel ist auch die Wertschätzung wirklicher Qualität und die Förderung der Freiheit der Wissenschaft zur Entdeckung von Neuland (siehe [14], [15] für weitergehende Diskussionen).

Zitieren wir zum Abschluss noch einmal aus der Schlusszusammenfassung des Dokuments der Akademien: „Evaluationen erfordern Bewertung durch anerkannte Experten höchster ethischer Integrität, mit Schwerpunkt auf intellektueller und wissenschaftlicher Leistung der zu Bewertenden. Bibliometrische Daten sind kein Ersatz für die Bewertung durch Experten, eine gute begründete inhaltliche Beurteilung ist essentiell. Die Nutzung von „Indikatoren" (metrics) kann wissenschaftlicher Kreativität und Originalität ernsthaften Schaden zufügen". Oder als leicht amerikanisiertes Kurzzitat:

„Scientist's Hirsch Index Tracking = S..."

LITERATUR

[1] M. Quack, über Preise der Bunsengesellschaft und Weiteres, *Bunsen-Magazin* 2015, **17**, 165-167

[2] M. Quack, Leitsätze für eine gute wissenschaftliche Publikationspraxis (mit einem Dokument der Académie des Sciences, der Leopoldina und der Royal Society), *Bunsen-Magazin* 2017, **19**, 41-43

[3] R. Cagan, The San Francisco Declaration on Research Assessment (DORA), 2013, Disease Models and Mechanisms, DOI 10:1242/dmm012955 and https://sfdora.org

[4] A. Molinié and G. Bodenhausen, „Bibliometrics as Weapons of Mass Citation", *Chimia* 2010, **64**, 78-89 (siehe auch [5] und [6] und *Bunsen-Magazin* 2010, **12**, 188-198)

[5] R. R. Ernst „The Follies of Citation Indices and Academic Ranking Lists. A Brief Commentary to Bibliometrics as Weapons of Citation" *Chimia* 2010, **64**, 90, Bunsen-Magazin 2010, **12**, 199-200

[6] A. Molinié, G. Bodenhausen, *Chimia*, 2011, **65**, 433-436

[7] W. Quan, B. Chen, F. Shu, „Publish or impoverish. An investigation of the monetary reward system of science in China 1999-2016", arXiv.org/abs/1707.01162

[8] A. Einstein, „Lens-like action of a star by the deviation of light in the gravitational field", *Science*, 1936, **84**, 506-507.

[9] A. Einstein, Sitzungsberichte Preuss. Akad. Wiss. 1918, Seiten 154-167 (siehe auch ebendort 1916, S. 688-690)

[10] D. Kneissl, H. Schwarz, „Grundlagenforschung braucht exzellente Wissenschaftler – und Freiräume. *Angew. Chemie* 2011, **133**, 12578-12579

[11] M. Quack, *Bunsen-Magazin* 2012, **14**, 181-189

[12] R. N. Zare, „Research Impact", *Current Science* 2012, **102**, 12

[13] R. R. Ernst, Die Verantwortung von Forschern: Eine europäische Sicht, *Angew. Chemie* 2003, **115**, 4572-4578

[14] M. Quack, „Über Autonomie und Freiheit der Wissenschaft: Mythen, Risiken und Chancen bei der Evaluation und Förderung der Naturwissenschaftlichen Grundlagenforschung, *Debatte* Heft **14**, 2015, 21-41 (Berlin Brandenburgische Akademie der Wissenschaften) und VSH Bulletin der Vereinigung schweizerischer Hochschuldozenten Heft 3/4 (2016), S. 61-69, ISSN 1663-9898

[15] M. Quack, Wie kommt das Neue in die Naturwissenschaft?, *Debatte* Heft **15** (2015), S. 29-58, BBAW, Berlin 2015

ANHANG

THE ROYAL SOCIETY

December 2017

Statement by three national academies (Académie des Sciences, Leopoldina and Royal Society) on good practice in the evaluation of researchers and research programmes

1. Introduction

The large increase in the size of the international scientific community, coupled with the desire to ensure the appropriate and efficient use of the substantial funding devoted to supporting scientific research, have understandably led to an increased emphasis on accountability and on the evaluation of both researchers, research activities and research projects (including recruitment, as well as the evaluation of grants and prizes). Given that there is a large diversity of procedures currently used in evaluations which have accumulated over time, it is now necessary to provide some guidelines for best practice in the evaluation of scientific research. Peer review, adhering to strict standards, is widely accepted as by far the best method for research evaluation. In this context, the present statement focuses on the evaluation of individual researchers.

Such an assessment by competent experts should be based on both written (journal articles, reviews, books, book chapters, patents, etc.) and other contributions and indicators of esteem (conference presentations, awards, public engagement activity, peer review activity, datasets shared, seminars, etc.). As a careful evaluation of scientific content and quality by experts is time consuming and costly, the number of evaluations should be limited and only undertaken when necessary, in particular for decisions on competitive academic appointments or funding large projects.

With the increase in the number of evaluations and the emergence of easily accessible electronic databases, the use of bibliometric measures has become an additional tool. However, there has been too much reliance on bibliometric indices and indicator-based tools as measures of performance by many evaluation committees and exercises, leading to the

danger of superficial, over-simplified and unreliable methods of evaluation. This bad practice involving the misuse of metrics has become a cause for serious concern.

Of particular concern are the widely used journal impact factors (IF) which are an estimate of the impact of the journal itself rather than the intrinsic scientific quality of a given article published within it – a point that has been made on several occasions and notably in the San Francisco Declaration[(1)]. Outstanding and original work can be found published in journals of low impact factor and the converse is also true. Nevertheless, the use of impact factors as a proxy for the quality of a publication is now common in many disciplines. There is growing concern that such "IF pressure" on authors has increased the incidence of bad practice in research and the 'gaming' of metrics over the past two decades, in particular in those disciplines that have over-emphasized impact factors. Also, the so–called 'altmetrics' – a new form of impact measure – while adding an important and hitherto overlooked dimension to the measurement of impact, suffers from some of the same weaknesses as the existing citation-based metrics.

There is a serious danger that undue emphasis on bibliometric indicators will not only fail to reflect correctly the quality of research, but may also hinder the appreciation of the work of excellent scientists outside the mainstream; it will also tend to promote those who follow current or fashionable research trends, rather than those whose work is highly novel and which might produce completely new directions of scientific research. Moreover, over-reliance on citations as a measure of quality may encourage the formation of aggregates of researchers (or "citation clubs") who boost each others citation metrics by mutual citation. It thus becomes important to concentrate on better methods of evaluation, which promote good and innovative scientific research.

2. Principles of good practice in the evaluation of researchers and research activities

Essential elements for the evaluation of researchers can be summarized as follows:

2.1. Selection of evaluation procedures and evaluators

Evaluators

Since the evaluation of research by peers is the essential process by which its quality and originality can be estimated, it is crucial to ensure that the evaluators themselves adhere to the highest standards and are leaders in their field. The selection of evaluators should be based on their scientific excellence and integrity. Their scientific achievements should be widely recognised and their curriculum vitae and research achievements should be easily accessible. Such an open process will ensure the credibility and transparency of the evaluations.

Evaluation processes

Since the number of excellent evaluators is limited, the number of evaluation processes should be reduced in order to avoid over-use of first-class evaluators. There is a concern that different agencies and institutions have carried out an excessive number of routine evaluations over the last decades, putting too much pressure on the best evaluators. First-rate evaluators are increasingly reluctant to commit to time-consuming and unproductive evaluation exercises. It is of great importance to reduce the number of evaluations and to confine them to the core issues of research that only peers are able to judge. Evaluators provide a "free resource" as part of their academic duty and this resource is over-exploited. Evaluating bodies must recognise that good evaluation is a limited and precious resource.

A page limit for submissions to all evaluation processes is needed. Excessively long submissions are counter-productive: evaluators need to be able to concentrate on the essentials, which is problematic with very lengthy submissions.

Rotation of evaluators is essential to avoid excessive or repeated influence from the same opinion leaders. The panel of experts should be adapted to reflect the diversity of disciplines or scientific domains. Although gender and geographical distribution will be factors in the selection of evaluating groups, excellence must remain the primary criterion.

2.2. Ethical guidelines and duties of evaluators

Evaluators should clearly declare possible conflicts of interest before the evaluation process. The confidentiality of expert reviews and of the discussions in the evaluation panel must be strictly respected to protect both the evaluators and the evaluated persons.

While reviewers have often learned the practice of evaluation by experience and self-teaching, this competence cannot be taken as given. Methods and approaches to evaluating and reviewing should become part of all researchers' competence as should the ethical principles involved. Evaluators should be made aware of the dangers of "unconscious bias". There should, as far as possible, be equivalent standards and procedures for all research disciplines.

The evaluation procedures must also include mechanisms to identify the cases of biased or otherwise inappropriate reviews and exclude them from consideration.

2.3. Evaluation criteria

Evaluations must be based under all circumstances on expert assessment of scientific content, quality and excellence. Publications that are identified by the authors as their most important work, including major articles and books, should receive particular attention in the evaluation. The simple number of publications should not be a dominant criterion.

Impact factors of journals should not be considered in evaluating research outputs. Bibliometric indicators such as the widely used H index or numbers of citations (per article

or per year) should only be interpreted by scientific experts able to put these values within the context of each scientific discipline. The source of these bibliometric indicators must be given and checks should be made to ensure their accuracy by comparison to rival sources of bibliometric information. The use of bibliometric indicators should only be considered as auxiliary information to supplement peer review, not a substitute for it.

The use of bibliometric indicators for early career scientists must in particular be avoided. Such use will tend to push scientists who are building their career into well-established/fashionable research fields, rather than encouraging them to tackle new scientific challenges.

For patents a clear distinction should be made between the stages of application, delivery and licensing.

Success in raising research grant funding should, where relevant, be only one and not the dominant factor in assessing research performance. The main criteria must be the quality, originality and importance of the scientific research.

3. Short summary of the main recommendations

Evaluation requires peer review by acknowledged experts working to the highest ethical standards and focusing on intellectual merits and scientific achievements. Bibliometric data cannot be used as a proxy for expert assessment. Well-founded judgment is essential. Over-emphasis on such metrics may seriously damage scientific creativity and originality. Expert peer review should be treated as a valuable resource.