frontiers in
COMPUTATIONAL NEUROSCIENCE

HYPOTHESIS AND THEORY ARTICLE
published: 10 November 2011
doi: 10.3389/fncom.2011.00048

Nine criteria for a measure of scientific output

Gabriel Kreiman[1,2]* **and John H. R. Maunsell**[3]

[1] Department of Ophthalmology, Children's Hospital, Harvard Medical School, Boston, MA, USA
[2] Department of Neurology, Children's Hospital, Harvard Medical School, Boston, MA, USA
[3] Department of Neurobiology, Harvard Medical School, Boston, MA, USA

Edited by:
Nikolaus Kriegeskorte, Medical Research Council Cognition and Brain Sciences Unit, UK

Reviewed by:
Yasser Roudi, Norwegian University of Science and Technology, Norway
Konrad Koerding, Northwestern University, USA

***Correspondence:**
Gabriel Kreiman, Department of Ophthalmology, Children's Hospital, Harvard Medical School, 1 Blackfan Circle, Karp 11217, Boston, MA 02115, USA.
e-mail: gabriel.kreiman@tch.harvard.edu

Scientific research produces new knowledge, technologies, and clinical treatments that can lead to enormous returns. Often, the path from basic research to new paradigms and direct impact on society takes time. Precise quantification of scientific output in the short-term is not an easy task but is critical for evaluating scientists, laboratories, departments, and institutions. While there have been attempts to quantifying scientific output, we argue that current methods are not ideal and suffer from solvable difficulties. Here we propose criteria that a metric should have to be considered a good index of scientific output. Specifically, we argue that such an index should be quantitative, based on robust data, rapidly updated and retrospective, presented with confidence intervals, normalized by number of contributors, career stage and discipline, impractical to manipulate, and focused on quality over quantity. Such an index should be validated through empirical testing. The purpose of quantitatively evaluating scientific output is not to replace careful, rigorous review by experts but rather to complement those efforts. Because it has the potential to greatly influence the efficiency of scientific research, we have a duty to reflect upon and implement novel and rigorous ways of evaluating scientific output. The criteria proposed here provide initial steps toward the systematic development and validation of a metric to evaluate scientific output.

Keywords: impact factors, peer review, productivity, scientific output, citation, bibliometric analysis, quality versus quantity, impact

INTRODUCTION

Productivity is the ratio of some output value to some input value. In some enterprises productivity can be measured with high precision. A factory can easily measure how many widgets are produced per man-hour of labor. Evaluating scientific productivity, however, is trickier. The input value for scientific productivity is tractable: it might be measured in terms of years of effort by a scientist, research team, department or program, or perhaps in terms of research dollars. It is the output value for scientific productivity that is problematic.

Scientific research produces new knowledge, some fraction of which can lead to enormous returns. In the long run, science evaluates itself. History has a particularly rigorous way of revealing the value of different scientific theories and efforts. Good science leads to novel ideas and changes the way we interpret physical phenomena and the world around us. Good science influences the direction of science itself, and the development of new technologies and social policies. Poor science leads to dead ends, either because it fails to advance understanding in useful ways or because it contains important errors. Poor science produces papers that can eventually feed the fireplace, or in a more modern and ecologically friendly version, the accumulation of electronic documents.

The process of science evaluating itself is slow. Meanwhile, we need more immediate ways of evaluating scientific output. Sorting out which scientists and research directions are currently providing the most useful output is a thorny problem, but it must be done. Scientists must be evaluated for hiring and promotion, and informed decisions need to be made about how to distribute research funding. The need for evaluation goes beyond the level of individuals. It is often important to evaluate the scientific output of groups of scientists such as laboratories, departments, centers, whole institutions, and perhaps even entire fields. Similarly, funding organizations and agencies need to evaluate the output from various initiatives and funding mechanisms.

Scientific output has traditionally been assessed using peer review in the form of evaluations from a handful of experts. Expert reviewers can evaluate the rigor, value and beauty of new findings, and gauge how they advance the field. Such peer-review constitutes an important approach to evaluating scientific output and it will continue to play a critical role in many forms of evaluation. However, peer review is limited by its subjective nature and the difficulty of obtaining comments from experts that are thorough and thoughtful, and whose comments can be compared across different evaluations. These limitations have driven institutions and agencies to seek more quantitative measures that can complement and sometimes extend thorough evaluation by peers.

In the absence of good quantitative measures of scientific output, many have settled for poor ones. For example, it is often assumed, explicitly, or implicitly, that a long list of publications indicates good output. Using the number of publications as a metric emphasizes quantity rather than quality, when it is the latter that is almost always the value of interest (Siegel and Baveye, 2010; Refinetti, 2011). In an attempt to measure something closer to quality, many turn to journal impact factors (Garfield, 2006). The misuse of journal impact factors in evaluating scientific output has been discussed many times (e.g., Hecht et al., 1998; Amin and Mabe, 2000; Skorka, 2003; Hirsch, 2005; Editors, 2006; Alberts et al., 2008; Castelnuovo, 2008; Petsko, 2008; Simons, 2008; Bollen et al., 2009; Dimitrov et al., 2010; Hughes et al., 2010 among many others). We will not repeat the problems with using the impact factors of *journals* to evaluate the output of *individual scientists* here, nor will we focus on the negative effects this use has had on the process of publishing scientific articles. Instead, we note that the persistent misuse of impact factors in the face of clear evidence of its inadequacies must reflect desperation for a quantitative measure of scientific output.

Many measures of scientific output have been devised or discussed. Because most scientific output takes the form of publication in peer-reviewed journals, these measures focus on articles and citations (Bollen et al., 2009). They include a broad range of approaches, such as total number of citations, journal impact factors (Garfield, 2006), *h*-factor (Hirsch, 2005), page ranks, article download statistics, and comments using social media (e.g., Mandavilli, 2011). While all these approaches have merit, we believe that no existing method captures all the criteria that are needed for a rigorous and comprehensive measure of scientific output. Here we discuss what we consider necessary (but not necessarily sufficient) criteria for a metric or index of scientific output. The goal of developing quantitative criteria to evaluate scientific output is not to replace examination by expert reviewers but rather to complement peer-review efforts. The criteria that we propose are aimed toward developing a quantitative metric that is appropriately normalized, emphasizes the quality of scientific output, and can be used for rigorous, reliable comparisons. We do not propose a specific measure, which should be based on extensive testing and comparison of candidate approaches, together with feedback from interested parties. Nevertheless, we believe that a discussion of properties that would make a suitable measure may help progress toward this goal.

We propose that a good index of scientific output will need to have nine characteristics.

DATA QUALITY AND PRESENTATION

QUANTITATIVE

Perhaps the most important requirement of a good measure of scientific output is that it be quantitative. The primary alternative, subjective ratings by experts will continue to be important for evaluations, but nevertheless suffers from some important limitations. Ratings by a handful of invited peers, as is normally used in hiring and promoting of scientists, provide ratings of undetermined precision. Moreover, the peers providing detailed comments on different job candidates or grant applications are typically non-overlapping, making it difficult to directly compare their comments.

A further problem with subjective comments is that they put considerable demands on reviewers' time. This makes it impractical to overcome uncertainties about comparisons between different reviewers by reaching out to a very large pool of reviewers for detailed comments. The alternative of getting brief comments from a very large pool of reviewers is also unlikely to work. Several initiatives provide frameworks for peer commentary from large sets of commenters. Most online journals provide rapid publication of comments from readers about specific articles (e.g., electronic responses for journals hosted by HighWire Press). However, few articles attract many comments, and most get none. The comments that are posted typically come from people with interest in the specific subject of the article, which means there is little overlap in the people commenting on articles in different journals. Even with comments from many peers, it remains unclear how a large set of subjective comments should be turned into a decision about scientific output.

BASED ON ROBUST DATA

Some ventures have sought to quantify peer commentary. For example, The Faculty of 1000 maintains a large editorial board for post-publication peer review of published articles, with numerical rating being given to each rated article. Taking another approach, WebmedCentral is a journal that publishes reviewers' comments and quantitative ratings along with published articles. However, only a small fraction of published articles are evaluated by systems like these, and many of these are rated by one or two evaluators, limiting the value of this approach as a comprehensive tool for evaluating scientific contributions. It is difficult to know how many evaluations would be needed to provide a precise evaluation of an article, but the number is clearly more than the few that are currently received for most articles. Additionally, it is difficult to assess the accuracy of the comments (should one also evaluate the comments?).

It seems very unlikely that a sufficiently broad and homogeneous set of evaluations could be obtained to achieve uniformly widespread quantitative treatment of most scientists while avoiding being dominated by people who are most vocal or who have the most free time (as opposed to people with the most expertise). There is also reason for concern that peer-rating systems could be subject to manipulation (see below). For these reasons, we believe that a reliable measure of scientific output should be based on hard data rather than subjective ratings.

One could imagine specific historical instances where subjective peer commentary could have been (and probably was) quite detrimental to scientific progress. Imagine Galileo's statement that the Earth moves or Darwin's Theory of Evolution being dismissed by Twitter-like commentators.

BASED ON DATA THAT ARE RAPIDLY UPDATED AND RETROSPECTIVE

While other sources might be useful and should not be excluded from consideration, the obvious choice for evaluation data is the citations of peer-reviewed articles. Publication of findings in peer-reviewed journals is the *sine qua non* for scientific progress, so the scientific literature is the natural place to look for a measure of scientific output. Article citations fulfill several important criteria. First, because every scientist must engage in scientific publication, a measure based on citations can be used to assess any scientist or group of scientists. Second, data on article citations are readily accessible and updated regularly, so that an index of output can be up-to-date. This may be particularly important for evaluating junior scientists, who have a short track record. Finally, publication data are available for a period that spans the lives of almost all working scientists, making it possible to track trends or monitor career trajectories. Historical data are particularly important for validating any measure of scientific output (see below), and would be impractical to obtain historical rankings using peer ratings or other subjective approaches. Because citations provide an objective, quantifiable, and available resource, different indices can be compared (see Validation below) and incremental improvements can be made based on evaluation of their relative merits.

Citations are not without weaknesses as a basis for measuring scientific output. While more-cited articles tend to correlate with important new findings, articles can also be cited more because they contain important errors. Review articles are generally cited more than original research articles, and books or chapters are generally cited less. Although articles are now identified by type in databases, how these factors should be weighted in determining an individual's contribution would need to be carefully addressed in constructing a metric. Additionally, there will be a lag between publication and citations due to the publishing process itself and due to the time required to carry out new experiments inspired by that publication.

Citations also overlook other important components of a scientist's contribution. Scientists mentor students and postdoctorals, teach classes and give lectures, organize workshops, courses and conferences, review manuscripts and grants, generate patents, lead clinical trials, contribute methods, algorithms and data to shared repositories and reach out to the public through journalists, books, or other efforts. For this reason, subjective evaluations by well-qualified experts are likely to remain an essential component of evaluating scientific output. Some aspects of the scientific output not involving publication might be quantified and incorporated into an index of output, but some are difficult to quantify. Because it is likely that a robust index of scientific output will depend to a large extent on citation data, in the following section we restrict our discussion to citations, but without intending to exclude other data that could contribute to an index (which might be multidimensional).

We acknowledge that there are practical issues that will need to be overcome to create even the simplest metric based on citations. In particular, to perform well it will be necessary for databases to assign a unique identifier to individual authors, without which it would be impossible to evaluate anyone with names like Smith, Martin, or Nguyen. However, efforts such as these should not be a substantial obstacle and some are already underway (e.g., Author ID by PubMed or ArXiv, see Enserink, 2009).

PRESENTED WITH DISTRIBUTIONS AND CONFIDENCE INTERVALS

An index of scientific output must be presented together with an appropriate distribution or confidence interval. Considering variation and confidence intervals is commonplace in most areas of scientific research. There is something deeply inappropriate about scientists using a measure of performance without considering its precision. A substantial component of the misuse of impact factor is the failure to consider its lack of precision (e.g., Dimitrov et al., 2010).

While the confidence intervals for an index of output for prolific senior investigators or large programs might be narrow, those for junior investigators will be appreciable because they have had less time to affect their field. Yet it is junior investigators who are most frequently evaluated for hiring or promotion. For example, when comparing different postdoctoral candidates for a junior faculty position, it would be desirable to know the distribution of values for a given index across a large population of individuals in the same field and at the same career stage so that differences among candidates can be evaluated in the context of this distribution. Routinely providing a confidence interval with an index of performance will reveal when individuals are statistically indistinguishable and reduce the chances of misuse.

NORMALIZATION AND FAIRNESS

NORMALIZED BY NUMBER OF CONTRIBUTORS

When evaluating the science reported in a manuscript, the quality and significance of the work are the main consideration, and the number of authors that contributed the findings is almost irrelevant. However, the situation differs when evaluating the contributions of individuals. Clearly, if a paper has only one author, that scientist deserves more credit for the work than if that author published the same paper with 10 other authors.

Defining an appropriate way to normalize for the number of contributors is not simple. Dividing credit equally among the authors is an attractive approach, but in most cases the first author listed has contributed more to an article than other individual authors. Similarly, in some disciplines the last place in the list is usually reserved for the senior investigator, and the relative credit due to a senior investigator is not well established.

Given the importance of authorship, it would not be unreasonable to require authors to explicitly assign to each author a quantitative fractional contribution. However, divvying up author credit quantitatively would not only be extremely difficult but would also probably lead to authorship disputes on a scale well beyond those that currently occur when only the order of authors must be decided. Nevertheless, some disciplines have already taken steps in this direction, with an increasing number of journals requiring explicit statements of how each author contributed to an article.

While it seems difficult to precisely quantify how different authors contribute to a given study, if such an approach came into practice, it might not take long before disciplines established standards for assigning appropriate credit for different types of contributions. Regardless of how normalization for the number of authors is done, one likely benefit of a widely used metric normalized in this way would be the rapid elimination of honorary authorship.

NORMALIZED BY DISCIPLINE

Scientists comprise overlapping but distinct communities that differ considerably in their size and publication habits. Publications in some disciplines include far more citations than others, either because the discipline is larger and produces more papers, or because it has a tradition of providing more comprehensive treatment of prior work (e.g., Jemec, 2001; Della Sala and Crawford, 2006; Bollen et al., 2009; Fersht, 2009). Other factors can affect the average number of citations in an article, such as journals that restrict the number of citations that an article may include.

A simple index based on how frequently an author is cited can make investigators working in a large field that is generous with citations appear more productive than one working in a smaller field where people save extensive references for review articles. For example, if two fields are equivalent except that articles in one field reference twice the number of articles as the other field, a simple measure based on citations could make scientists in the first field appear on average to be twice as productive as those in the second. To have maximal value, an index of output based on citations should normalize for differences in the way that citations are used in different fields (including number of people in the field, etc.). Ideally, a measure would reflect an individual's relative contribution *within* his or her field. It will be challenging to produce a method to normalize for such differences between disciplines in a rigorous and automatic way. Comprehensive treatment of this issue will require simulation and experimentation. Here, we will briefly mention potential approaches to illustrate a class of solutions.

There is a well-developed field of defining areas of science based on whether pairs of authors are cited in the same articles (author co-citation analysis; Griffith et al., 1986). More recently, these methods have been extended by automated rating of text similarity between articles (e.g., Greene et al., 2009). Methods like these might be adopted to define a community for any given scientist. With this approach, an investigator might self-define their community based on the literature that they consider most relevant, as reflected by the articles they cite in their own articles. For a robust definition that could not be easily manipulated (see below), an iterative process that used articles that cite cited articles, or articles that are cited by cited articles, would probably be needed. While it is difficult to anticipate what definition of a scientist's community might be most effective, one benefit of using objective, accessible data is that alternative definitions can be tested and refined.

Once a community of articles has been defined for an investigator, the fraction of all the citations in those articles that refer to the investigator would give a measure of the investigator's impact within that field. This might provide a much more valuable and interpretable measure than raw counts of numbers of papers or number of citations. It is conceivable that this type of analysis could also permit deeper insights. For example, it might reveal investigators who were widely cited within multiple communities, who were playing a bridging role.

NORMALIZED FOR CAREER STAGE

A measure that incorporated the properties discussed so far would allow a meaningful assessment of an individual's contribution to science. It would, however, rate senior investigators as more influential than junior investigators. This is a property of many existing measures, such as total number of citations or h-index. For some purposes this is appropriate; investigators are frequently compared against others at a similar stage of their careers, and senior scientists generally have contributed more than junior scientists. However, for some decisions, such as judging which investigators are most productive per unit time, an adjustment for seniority is needed. Additionally, it might be revealing for a search committee to compare candidates for an Assistant Professor position with well-known senior investigators when they entered the rank of Assistant Professor.

This type of normalization for stage of career would be difficult to achieve for several reasons. The explosive growth in the number of journals and scientists will make precise normalization difficult. Additionally, data for when individuals entered particular stages (postdoctoral, Assistant Professor, Associate Professor, Full Professor) are not widely available. A workable approximation might be possible based on the time since an author's first (or first n) papers were published. Because the size of different disciplines changes with time, and the rate at which articles are cited does not remain constant, these trends would need to be compensated in making comparisons over time.

A related issue is the effect of time itself on citation rates. An earlier publication has had more time to be cited (yet scientists tend to cite more recent work). In some sense, a publication from the year 2000 with 100 citations is less notable than a publication from the year 2010 with 100 citations. A simple way to address this is to compute the number of citations per year (yet we note that this involves arguable assumptions of stationarity in citation rates).

FOSTERING GREAT SCIENCE

IMPRACTICAL TO MANIPULATE

If a metric can be manipulated, such that it can be changed through actions that are relatively easy compared to those that it is supposed to measure, people will undoubtedly exploit that weakness. Given an index that is based on an open algorithm (and the algorithm should be open, computable and readily available), it is inevitable that scientists whose livelihoods are affected by that index will come up with ingenious ways to game the system. A good index should be impractical to game so that it encourages scientists to do good science rather than working on tactics that distort the measure.

It is for this reason that measures such as the number of times an article is downloaded cannot be used. That approach would invite the generation of an industry that would surreptitiously download specific articles many times for a fee. For the same reason, a post-publication peer-review measure that depended on evaluations from small numbers of evaluators cannot be robust when careers are at stake.

A measure that is based on the number of times an author's articles are cited should be relatively secure from gaming, assuming that the neighborhood of articles used to normalize by discipline is sufficiently large. Even a moderate-sized cartel of scientists who agreed to cite each other gratuitously would have little impact on their metrics unless their articles were so poorly cited that any manipulation would still leave them uncompetitive. Nevertheless, it seems likely that a measure based on citations should ignore self-citations and perhaps eliminate or discount citations from recent co-authors (Sala and Brooks, 2008).

One would hope that a key motivation for scientific inquiry is, as Feynman put it, "the pleasure of finding things out." Yet, any metric to evaluate scientific output establishes a certain incentive structure in the research efforts. To some extent, this is unavoidable. Ideally, the incentive structure imposed by a good metric should promote great science as opposed to incentive structures that reward (even financially in some cases) merely publishing an article in specific journals or publishing a certain number of articles. A good metric might encourage collaborative efforts, interdisciplinary efforts, and innovative approaches. It would be important to continuously monitor and evaluate the effects of incentive structures imposed by any metric to ensure that they do not discourage important scientific efforts including interdisciplinary research, collaborations, adequate training, and mentoring of students and others.

FOCUSED ON QUALITY OVER QUANTITY

Most existing metrics show a monotonic dependence on the number of publications. In other words, there are no "negative" citations (but perhaps there should be!). This monotonicity can promote quantity rather than quality. Consider the following example (real numbers but fictitious names). We compare authors Joe Doe and Jane Smith who work in the same research field. Both published his or her first scientific article 12 years ago and the most recent publication from each author was in 2011. Joe has published 45 manuscripts, which have been cited a total of 591 times (mean = 13.1 citations per article, median = 6 citations per article). Jane has published 14 manuscripts, which have been cited 1782 times (mean = 127.3 citations per article median = 57 citations per article). We argue that Jane's work is more impactful in spite of the fact that her colleague has published three times more manuscripts in the same period of time. The process of publishing a manuscript has a cost in itself including the time required for the authors to do the research and report the results, the time spent by editors, reviewers, and readers to evaluate the manuscript.

In addressing this issue, care must be taken to avoid a measure that discourages scientists from reporting solid, but apparently unexciting, results. For example, penalizing the publication of possibly uninteresting manuscripts by using the average number of citations per article would be inappropriate because it would discourage the publication of any results of below-average interest. The *h*-index (and variants) constitutes an interesting attempt to emphasize quality (Hirsch, 2005). An extension of this notion would be to apply a threshold to the number of citations: publications that do not achieve a certain minimum number of citations would not count toward the overall measure of output. This threshold would have to be defined empirically and may itself be field-dependent. This may help encourage scientists to devote more time thinking about and creating excellence rather than wasting everyone's time with publications that few consider valuable.

VALIDATION

Given a metric, we must be able to ask how good it is. Intuitively, one could compare different metrics by selecting the one that provides a better assessment of excellence in scientific output. The argument, however, appears circular because it seems that we need to have *a priori* information about excellence to compare different possible metrics. It could be argued that the scientific community will be able to evaluate whether a metric is good or not by assessing whether it correlates well with intuitive judgments about what constitutes good science and innovative scientists. While this is probably correct to some extent, this procedure has the potential to draw the problem back to subjective measures.

To circumvent these difficulties, one could attempt to develop quantitative criteria to evaluate the metrics themselves. One possibility is to compare each proposed quantitative metric against independent evaluations of scientific output (which may not be quantitative or readily available for every scientist). For example, Hirsch (2005) attempted to validate the *h*-index by considering Nobel laureates and showing that they typically show a relatively large *h*-index. In general, one would like to observe that the metric correlates with expert evaluations across a broad range of individuals with different degrees of productivity. While this approach seems intuitive and straightforward it suffers from bringing the problem back to subjective criteria.

An alternative may be to consider historical data. A good metric could provide *predictive* value. Imagine a set of scientists and their corresponding productivity metric values evaluated in the year 2011. We can ask how well we can predict the productivity metric values in 2011 from their corresponding values in the year 2000 or 1990. Under the assumption that the scientific productivity of a given cohort is *approximately* stationary, we expect that a useful metric would show a high degree of prediction power whereas a poor metric will not. Of course, many factors influence scientific productivity over time for a given individual and these would be only correlative and probabilistic inferences. Yet, the predictive value of a given metric could help establish a quantitative validation process.

Given the importance of evaluating scientific output, the potential for a plethora of metrics and the high-dimensional parameter landscape involved, it seems worth further examining and developing different and more sophisticated ways of validating these metrics. One could consider measures of scientific influence based on the spread of citations, the number of successful trainees, etc., and compare these to different proposed metrics. Ultimately, these are empirical questions that should be evaluated with the same rigor applied to other scientific endeavors.

DISCUSSION

We describe above nine criteria that, we hope, might lead to a better way of evaluating scientific output. The development of an evaluation algorithm and metric that capture these properties is not intended to eliminate other forms of peer evaluation. Subjective peer review is valuable (both pre-publication and post-publication) despite its multiple pitfalls and occasional failures, and a combination of different assessments will provide more information than any one alone.

A metric that captured the properties discussed above could provide many benefits. It might encourage better publishing practices by discouraging publication of a large number of uneventful reports or reducing the emphasis on publishing in journals with high impact factors. By highlighting the scientific contributions of individuals within a field it might restore a more appropriate premium: providing important results that other scientists feel compelled to read, think about, act upon, and cite. Placing emphasis on how often other scientists cite work may have other beneficial effects. A long CV with many least-publishable papers would quickly become visibly inferior to a shorter one with fewer but more influential papers. As mentioned above, there may be other benefits including correcting authorship practices, accurate evaluation across disciplines, and it may even help students choose a laboratory or institution for graduate studies or postdoctoral research.

In addition to evaluating the current value of a productivity metric, it may be of interest to compute the rate of change in this metric. This might help highlight individuals, laboratories, departments, or institutions that have recently excelled. Rates should also be normalized and presented alongside distributions as discussed above for the metric itself.

Although we have cast the discussion in terms of a single metric, an index of output does not need to be scalar. No single value can capture the complexities involved in scientific output. Different aspects of an investigator's contributions may require different indices. Additionally, evaluating a research group, a research center, or a department may be distinct from evaluating an individual and require somewhat different metrics (e.g., Hughes et al., 2010), but once suitable measures of output are available, productivity can be evaluated in terms of either years of effort, number of people involved, research funding, and other relevant parameters.

No calculation can take the place of a thoughtful evaluation by competent peers, and even an index that is precise and accurate can be abused. Evaluators might blindly apply an index without actually assessing papers, recommendations, and other material. Evaluators might also ignore confidence intervals and try to make unjustified distinctions between the performance of individuals or programs with different, but statistically indistinguishable, metrics.

Given current technologies, the state of information science, and the wealth of data on authors, publications and citations, useful quantification of the scientific output of individuals should be attainable. While we have avoided the challenge of defining and validating specific algorithms, there is little doubt that a superior metric could be produced. Given how much is at stake in decisions about how to allocate research support, there is no excuse for failing to try to provide a measure that could end the misdirected use of impact factor, download statistics, or similar misleading criteria for judging the contributions of individuals. While the newly developed metrics may show some degree of correlation with existing ones, we have to develop indices that are question-specific (e.g., how do we evaluate a given scientist?) as opposed to using generic indices developed for other purposes (e.g., how do we evaluate a certain web site or journal?). Because it has the potential to greatly influence the efficiency of scientific research, we have a duty to reflect upon and eventually implement novel and rigorous ways of evaluating scientific output.

ACKNOWLEDGMENTS

We thank John Hogenesch, Martin Hemberg, Marlene Cohen, and Douglas Ruff for comments and discussions.

REFERENCES

Alberts, B., Hanson, B., and Kelner, K. L. (2008). Reviewing peer review. *Science* 321, 15.

Amin, M., and Mabe, M. (2000). Impact factors: use and abuse. *Perspect. Publ.* 1, 1–6.

Bollen, J., Van de Sompel, H., Hagberg, A., and Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE* 4, e6022. doi:10.1371/journal.pone.0006022

Castelnuovo, G. (2008). Ditching impact factors: time for the single researcher impact factor. *BMJ* 336, 789.

Della Sala, S., and Crawford, J. R. (2006). Impact factor as we know it handicaps neuropsychology and neuropsychologists. *Cortex* 42, 1–2.

Dimitrov, J. D., Kaveri, S. V., and Bayry, J. (2010). Metrics: journal's impact factor skewed by a single paper. *Nature* 466, 179.

Editors. (2006). The impact factor game. It is time to find a better way to assess the scientific literature. *PLoS Med.* 3, e291. doi:10.1371/journal.pmed.0030291

Enserink, M. (2009). Scientific publishing. Are you ready to become a number? *Science* 323, 1662–1664.

Fersht, A. (2009). The most influential journals: impact factor and Eigenfactor. *Proc. Natl. Acad. Sci. U.S.A.* 106, 6883–6884.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA* 295, 90–93.

Greene, D., Freyne, J., Smyth, B., and Cunningham, P. (2009). *An Analysis of Current Trends in CBR Research Using Multi-View Clustering.* Technical Report UCD-CSI-2009-03. Dublin: University College Dublin.

Griffith, B. C., White, H. D., Drott, M. C., and Saye, J. D. (1986). Tests of methods for evaluating bibliographic databases: an analysis of the National Library of Medicine's handling of literatures in the medical behavioral sciences. *J. Am. Soc. Inf. Sci.* 37, 261–270.

Hecht, F., Hecht, B. K., and Sandberg, A. A. (1998). The journal "impact factor": a misnamed, misleading, misused measure. *Cancer Genet. Cytogenet.* 104, 77–81.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U.S.A.* 102, 16569–16572.

Hughes, M. E., Peeler, J., and Hogenesch, J. B. (2010). Network dynamics to evaluate performance of an academic institution. *Sci. Transl. Med.* 2, 53ps49.

Jemec, G. B. (2001). Impact factor to assess academic output. *Lancet* 358, 1373.

Mandavilli, A. (2011). Peer review: trial by Twitter. *Nature* 469, 286–287.

Petsko, G. A. (2008). Having an impact (factor). *Genome Biol.* 9, 107.

Refinetti, R. (2011). Publish and flourish. *Science* 331, 29.

Sala, S. D., and Brooks, J. (2008). Multi-authors' self-citation: a further impact factor bias? *Cortex* 44, 1139–1145.

Siegel, D., and Baveye, P. (2010). Battling the paper glut. *Science* 329, 1466.

Simons, K. (2008). The misused impact factor. *Science* 322, 165.

Skorka, P. (2003). How do impact factors relate to the real world? *Nature* 425, 661.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 23 June 2011; accepted: 25 October 2011; published online: 10 November 2011.

*Citation: Kreiman G and Maunsell JHR (2011) Nine criteria for a measure of scientific output. Front. Comput. Neurosci. **5**:48. doi: 10.3389/fncom.2011.00048*