Evaluating how we evaluate

Ronald D. Vale
Department of Cellular and Molecular Pharmacology and the Howard Hughes Medical Institute, University of California, San Francisco, San Francisco, CA 94158

ABSTRACT **Evaluation of scientific work underlies the process of career advancement in academic science, with publications being a fundamental metric. Many aspects of the evaluation process for grants and promotions are deeply ingrained in institutions and funding agencies and have been altered very little in the past several decades, despite substantial changes that have taken place in the scientific work force, the funding landscape, and the way that science is being conducted. This article examines how scientific productivity is being evaluated, what it is rewarding, where it falls short, and why richer information than a standard curriculum vitae/biosketch might provide a more accurate picture of scientific and educational contributions. The article also explores how the evaluation process exerts a profound influence on many aspects of the scientific enterprise, including the training of new scientists, the way in which grant resources are distributed, the manner in which new knowledge is published, and the culture of science itself.**

Monitoring Editor
David G. Drubin
University of California, Berkeley

Received: Jun 29, 2012
Accepted: Jul 5, 2012

The scientific profession is fundamentally a meritocracy. As part of this meritocracy, our scientific work is constantly scrutinized through "peer review," a system that is solid and arguably adopts higher standards of fairness and rigor than those of many other occupations. Manuscripts are evaluated for publication by reviewers and journal editors, and scientists vie for precious real estate in what are perceived to be the prime journals. Published papers, in turn, are the most important metrics in evaluating grant applications and promotions.

While our evaluation process is based on the sound principle of peer review, the profession is facing new challenges of tightened research budgets in many countries and explosive growth in the number of scientists during the past two decades. Thus we are at a juncture at which it is reasonable to ask how well our evaluation systems for papers, grants, and promotions are working. Are they adapting to changes in science, new publication options, and new career structures? Are they producing and rewarding the best possible science and meeting the needs of young scientists? The goal of this article is to discuss these issues.

DOI: 10.1091/mbc.E12-06-0490

Address correspondence to: Ronald D. Vale (vale@cmp.ucsf.edu).

Abbreviations used: CV, curriculum vitae; HHMI, Howard Hughes Medical Institute; NIH, National Institutes of Health; PI, principal investigator.

ARE WE EVALUATING SCIENTIFIC QUALITY OR OUTSOURCING THIS RESPONSIBILITY TO JOURNALS?

"Let's try for *Science*, *Nature*, or *Cell*!" exclaim a student/postdoc and his/her advisor. These journals reach a wide audience, as many scientists frequently scan their tables of contents. However, scanning tables of contents has become less important now with the availability of search engines such as PubMed than it was in the past, when journals were retrieved one at a time from the shelves of a library. Thus it is somewhat counterintuitive that the three main journals have remained very powerful, when print subscriptions are in decline and most journals can be accessed electronically. The primary reason driving the current frenzied submission rate to these journals is the opportunity for career advancement. Publications in these journals are golden eggs in a curriculum vitae (CV) that can significantly enhance chances for getting jobs and grants.

In a meritocracy, evaluation of productivity is necessary, and judgment of quality must come into play. But have we dug ourselves into too deep of a hole by relying so heavily on journals and their associated impact factors for making decisions about quality? Have we "outsourced" too much of our responsibility in peer evaluation to journals?

A rationale for adopting a journal hierarchy as a proxy for quality is that top journals receive many papers; in partnership with scientific reviewers, they invest considerable energy in sorting through submissions to identify the "best science." While this may seem like a perfect Darwinian selection system, we also are all aware of its flaws (Simons, 2008; Johnston, 2009). Not infrequently, a paper in a "top journal" fades from sight after publication, while the subsequent impact of a paper in a "lesser journal" increases. Journals also

look for particularly newsworthy content to enhance their image (which they have the right to do) and not always for the best science. Given the large numbers of submissions, there also is a tendency to accept papers that have a clean bill of health from three or four reviewers, which is not necessarily a metric of outstanding science. The ultimate decision makers also are the journal editors, not the scientists who write the reviews. Thus this peer review system is heavily filtered in a nontransparent (or at least translucent) process that incorporates the goals of a journal. Furthermore, many scientists do not want to waste time on the "journal game," prefer open-source journals, or seek more page space for their published work. Thus many outstanding studies are never subjected to the "top journal litmus test" in the first place.

There is disgruntlement in our scientific community about the growing emphasis of the *where*, rather than the *what*, in evaluating publications. This emphasis is creating more submissions, as a paper is often serially submitted, initially reaching for the top and, if rejected, moving down the journal food chain until it finds a home. This wastes enormous time that could be spent on doing science and creates anxiety among students and postdocs. However, I would argue that the fault does not lie with journal editors and their staff; their job is to make their journals successful. It is our job as a scientific community to evaluate published scientific work. We have created the predicament in which we find ourselves.

If the ball is in the court of the scientific community, why have we clung so tightly to and even reinforced the journal hierarchy? Not uncommonly, scientists who complain about the system succumb to it when it is their turn to write/present a peer review evaluation. Scientists themselves have become seduced by the sparkle of a high-profile paper on a CV. With so many papers and a shortage of time for reading and understanding them, counting high-profile papers on a CV is an easy solution for a scientist with a busy schedule. Reducing complex science to easy impact factors also provides tools for administrators who do not themselves understand the science.

What can be done to dig ourselves out of this rut? The first step is recognizing that peer evaluation is *our* responsibility. In evaluating qualifications for a grant, a job, or a promotion, it is too simplistic to think that judgment has already been rendered by prior competition for the most prized journal pages. Second, our scientific community might do well to reassert the value of publishing outstanding science in specialty journals. The phrase "better fit for a specialty journal" has become an uncomplimentary, lethal blow in the review process. But this view was not held by previous generations of scientists. While *Science* and *Nature* have been *the* places to publish important and provocative short communications for more than a century, prior generations of scientists often chose to publish their more complete, but still high-impact, studies in journals such as the *Journal of General Physiology,* the *Journal of Biological Chemistry,* and the *Journal of Cell Biology.* More recently, excellent new journals (such as *Molecular Biology of the Cell* and *PLoS*) have been added as publication possibilities. However, online supplemental material in *Science* and *Nature* (which in reality few read) also has contributed to the lower stature of the longer-format specialty journals, since 5 years of work can now be contained in a 2000-word "print" article along with the larger reservoir of space available as online material. Third, expanding the group of broad-interest, highly ranked journals beyond the present holy trinity might take some of the pressure off the system (the new *eLife* journal being launched by the Howard Hughes Medical Institute [HHMI], the Wellcome Trust, and the Max Planck Society will hopefully help in this expansion).

Most importantly, the merit of scientific work must be assessed after publication, rather than solely during the journal review process. We do not have a good general scheme for achieving this, and efforts such as Faculty of 1000 have not been influential in affecting evaluations. However, rather than waiting for new schemes, scientists who conduct peer review need to make sufficient effort to assess and articulate the value of scientific studies. We need to restrain ourselves from just using journal names as primary evidence of merit. For example, it is not uncommon for a grant discussion or a promotion letter to say, "In the past five years, the principal investigator (PI) has published six papers, two of which were published in *Cell*." Chances are that the work is excellent and the PI highly productive, but we need to explain the science first and kick the habit of resting one's argument of scientific productivity by invoking the name of a high-profile journal, as if this is all the information needed.

NUMBERS OF PUBLICATIONS—HOW MANY IS ENOUGH?

In addition to journal impact factors, the number of publications generated by a scientist is used as evidence of productivity. However, the time required to publish a paper is increasing, and evaluating committees cannot be frozen in time with expectations of publication numbers from the past.

To illustrate this point, compare this week's *Nature* with an issue published 30 years ago. The Letters are about the same length and have three or four figures. However, a typical Letter in this week's issue will likely have four figures packed with multiple panels and accompanied by 10 additional multi-panel supplemental figures. Hidden from view is the fact that the authors had to spend 6 months performing experiments and rewriting the paper in response to 25 comments from four reviewers (Raff *et al.*, 2008). In contrast, the typical 1982 Letter also contained excellent science, but was usually less of a magnum opus, had no online supplementary material, and most likely was not subjected to a daunting and repetitive review process.

Science is not harder in 2012 than it was in 1982; in many ways, it has become easier with the many tools available. However, in my opinion, the experimental threshold for publication has escalated in the past 30 years. We have now reached a point at which publishing an outstanding study may require approximately 4 years of work to see it through to the end. There is increasing concern about the steady increase over the past two decades in the age at which young scientists obtain a job and become independent. However, this should come as no surprise, as it follows from the combined pressures to produce a magnum opus as well as additional papers. Academic institutions tend to be conservative in hiring. Rather than viewing a postdoc position as an apprenticeship and making job offers on the promise of talent, most institutions offer jobs to postdocs who offer a handsome dowry of papers, which increases their likelihood of subsequent grant funding, the lifeblood of every institution. Amassing this dowry of papers is taking a longer and longer period of time.

In my opinion, expecting two or three publications from a graduate student or a postdoc is not realistic, especially if we (as a profession) are serious about our intent to graduate students sooner and have postdocs start their independent careers earlier in life. Furthermore, institutional expectations for more publications can create pressures that run counter to the goal of producing more complete, interesting, and high-quality scientific studies.

MOVING BEYOND THE CV: EVALUATING "QUALITY VERSUS QUANTITY" AND "SPECIFIC ACTIVITY VERSUS TOTAL OUTPUT"

Past productivity weighs heavily in the evaluation of grants and promotions. The evidence for past productivity is usually the chronological list of numbered publications in a CV or the more

abbreviated National Institutes of Health (NIH) "biosketch." However, it might be worthwhile to consider other formats for listing publications that might provide additional information.

To focus on outstanding science versus simple numbers of publications in a CV, HHMI asks its investigators to list their five most important publications since their last review, along with brief explanations of why each paper is significant. Writing a few sentences about a paper also draws attention to the science created and impact generated and not just the journal name. This format requires one page and could be adopted in many kinds of evaluation settings.

To better compare the output from large and small laboratories, knowing the "denominator" (number of people working in the lab) in addition to the "numerator" (the number and, more importantly, quality of papers produced) is useful, since it provides a measure of "specific activity." This metric provides insight into the both the productivity and mentoring environment of the laboratory. Both the numerator and denominator, however, are complex measures and cannot be reduced to a simple division and numeric output. As articulated earlier, the numerator is an assessment of impact, not just a scorecard of the total number of papers produced. The denominator also is not a simple measure, since workers stay for various lengths of time. But as one example, one could list each researcher who had been in the laboratory for three or more years along with his or her publications (or less than 3 years as a bonus if that researcher was fortunate enough to have a publication in that time). Bringing the denominator into play may encourage a PI to be more concerned about who is accepted into a lab and how that person is mentored, which might benefit scientific training overall, as will be discussed later.

In summary, scientists are "data-driven" in their work. But when it comes to evaluating productivity, we operate with relatively little data at our disposal through standard CVs. Attempts have been made to provide more data in the form of the "h-index" (Ball, 2007) and "Z factors," but I think that such metrics worsen rather than improve our evaluation ability by encouraging more superficial review. Reducing performance to a few numbers makes sense for a baseball pitcher or a hockey goalie, but not for a scientist. Science is more complex, and peer evaluation involves good judgment of scientific and educational output.

EVALUATING BIG AND SMALL LABORATORIES

Bruce Alberts wrote a compelling essay in 1985 on the virtue of small laboratories and the frequent inefficiency of large laboratories; it is a still a good read today (Alberts, 1985). This essay drew attention and, in part, resulted in Alberts being selected to chair a National Research Council committee to evaluate whether to launch a massive effort to sequence the human genome, the very antithesis of a small lab. However, Alberts and his committee decided to endorse this large-scale project, foreseeing the opportunity and the exceptional potential impact (Olson, 1995).

This example illustrates that optimizing lab size is complex, involving many factors, and goal-dependent. Many small laboratories make remarkable discoveries that are the envy of much larger labs (indeed many Nobel Prize discoveries in the life sciences emerged from small-sized labs). Many large laboratories are highly productive by any size-normalized metric and use their combined resources in synergistic ways not easily achieved by a smaller laboratory. Some projects (e.g., the human genome) exceed what any single lab can do. Indeed, the mixture of small and large laboratories has been successful historically for covering the wide breadth of the life sciences, making discoveries and translating discoveries into practical outcomes that benefit society.

However, in an era in which PIs with a single NIH R01 grant are struggling to retain their grant support, and junior faculty are receiving their first R01 later in life (Ph.D. scientists now receive their first R01 at around age 42 [Matthews *et al.*, 2011]), it is more important than ever to ensure that larger laboratories and large projects are scrutinized. Creating an arbitrary ceiling for lab size is not the answer; if a PI can scale-up an operation with commensurate productivity and good mentoring of trainees, then such growth should be permitted and indeed encouraged. However, bigger does not necessarily mean better, and laboratories of different sizes should be evaluated for their productivity on a level playing field. In this regard, some reform may be in order.

In evaluating NIH grants, study sections are asked to focus on a proposal rather than an overall laboratory package; they do not evaluate productivity per dollar invested for laboratories that have multiple grants. A PI with a larger laboratory will generate more papers for his or her biosketches, a subset of which will land in high-profile journals. There are also no stringent guidelines that preclude investigators from "double-dipping" by listing a particular publication as evidence of past productivity for more than one grant. Thus a biosketch leaves a lot of guesswork in assessing the amount of money and personnel that goes into the evidence provided for past productivity. A larger lab also has a higher probability of generating papers or preliminary data in a new area, which can help seed a new R01 proposal. Some R01s from a large lab may not be resubmitted at the end of their funding period due to lack of progress, but they can be replaced by "fresh" R01s. Thus a larger lab will have a greater "buffer capacity" than a small laboratory supported by just one R01. Some entrepreneurial investigators establish a large amount of funding by serially submitting grants, each of which may look attractive based on the preliminary data and biosketch of the investigator. However, the net productivity of the entire laboratory might look less attractive if all of the grants were considered collectively.

With ample funding, the entrepreneurial system of building up support from separately evaluated R01s has worked well and produced good results. But during lean times, both maximizing productivity and supporting many investigators become important considerations. Large labs cannot simply be propagated into the future based on PI seniority or the "culture" of a particular field or institution.

How does one establish a level playing field for labs? While taboo at the moment, it may be time to allow study sections to evaluate the overall resources and productivity of a laboratory. Funding of multiple grants is taken into consideration by the NIH administration, but the output and discussion is not transparent or part of the peer review process. Changes to the biosketch that highlight "specific activity," as described earlier, might also be helpful. Furthermore, the NIH might convene a study to evaluate the circumstances in which large and small labs are particularly productive or generate unique outcomes in different fields of biology and biomedical research. Such information might help granting agencies to understand how best to optimize their resources.

A more radical change to the NIH system (and admittedly unlikely to happen!) would be to provide a single grant to a laboratory, similar to the way HHMI funds its investigators, rather than having PIs write multiple R01s. During an HHMI evaluation for funding, past work and future trajectory of the entire laboratory is summarized in a 3000-word document; all of the cards are on the table, making evaluation relatively straightforward. Applying this model to the NIH, young investigators could start with one module (e.g., $250,000/year, preferably fully funded to

avoid immediate pressure of obtaining more support). Later, if the lab performs well and has an exciting future plan, the grant could be renewed with an additional module of support; if an exceptional breakthrough is made, the PI could apply for early renewal with extra funding. Successful senior NIH investigators could still run larger labs supported by several modules. However, scientists who have been less productive and/or whose future work is less compelling may lose a module of support upon peer review. Modules, however, should not correspond to specific aims, which, if truly outstanding science is being done, should change and adapt over a 4- to 5-year grant period. Rather, the focus of review should be the overall potential of the lab (incorporating quality of past work and the sum of future directions) and not about study sections micromanaging funding modules by cutting a specific aim. If the performance of laboratories is evaluated, PIs will spend less time writing grants, the NIH will review fewer grants, and there will be less "gamesmanship" inherent in the process of writing multiple, independently reviewed grants to build up NIH support. Naturally, such a system would not preclude other types of funding schemes that serve different purposes, such as collaborative or technology development grants.

IS THE EVALUATION PROCESS DRIVING AN UNSUSTAINABLE WORKFORCE MODEL?

At a time when more laboratories need to be funded, our present evaluation process is fueling the fire to increase lab size. If institutions and granting agencies evaluate based on the number of publications and high-profile publications, then increasing the number of workers constitutes a viable strategy for a PI to meet these expectations. From my observations over the past two decades of being an academic scientist, the demand to publish more papers (each harder to produce) and high-profile papers (harder to come by) for tenure and grant support is driving junior faculty to escalate the size of their laboratories too quickly. Many junior faculty at major research institutions have more than 10 lab workers within 4 years. Indeed, a large lab in itself is perceived as a sign of success. Driven to achieve tenure and recognition in his or her field, the once "star" experimental postdoc is quickly reduced to an administrative PI trying to acquire sufficient grants to support a growing number of lab workers. This is a pity, since the young PI is most likely the most skilled and productive researcher in his or her own lab, and should have more time and freedom to pursue research. Many junior faculty (and indeed senior faculty) would be comfortable with and even prefer a smaller lab, if they were not penalized by the evaluation system for operating in such a style. Many European and some U.S. institutes (e.g., Janelia Farms Research Campus, HHMI) encourage and support a small junior leader model in their funding and evaluation mechanisms, but this is not true of most U.S. institutions, whose business model is based on bringing in as much grant support as possible and rewarding grant procurement in the evaluation process.

If junior faculty grow their labs quickly to meet evaluation expectations, then the graduate students and postdocs in their labs will emulate this model when they become junior faculty. Thus our evaluation system may be building up to create a perfect storm—a rapidly escalating, unsustainable workforce model that will dishearten young people who are on the doorstep. Exponential growth is glorious when there are plenty of nutrients in the broth. However, the research budget of the United States may be entering a "stationary phase" of research funding, which will warrant greater stewardship of the scientific workforce and education systems.

EVALUATING TRAINING OUTCOMES

The success of the scientific enterprise is dependent on its workforce. Because most PIs do not themselves work in the lab, their productivity is largely driven by their graduate students and postdocs. A natural symbiosis exists between scientific productivity and education of graduate students and postdocs, since completing studies and writing papers is a major part of learning the scientific trade. The synergy becomes precarious when the evaluation process becomes skewed toward demanding and rewarding production of papers at the expense of good mentoring.

Looking at training outcomes as well as the number of scientific papers produced may restore the balance between the goals of productivity and education. Successful training outcomes, not just placement in tenure-track academic careers, should be considered (Sauermann and Roach, 2012). In the current R01 application, a student or postdoc is listed as a "worker" who will produce the results described in the specific aims. Although not explicitly written in the grant instructions or budget justification, the NIH surely is providing this salary support to train these young people as scientists, not just as "workers" executing the aims of the grant. Postdoctoral fellowship–granting agencies often ask for outcomes (e.g., job placements) of past trainees. Thus it seems reasonable to request a mentoring plan within the R01 and to have study sections evaluate outcomes of recent past trainees, at least in cases in which support is requested for graduate student or postdoc salaries. Successful job outcomes also must be considered more broadly, with academic PIs representing just one part of a spectrum of diverse careers resulting from our training programs.

EVALUATING CONTRIBUTIONS IN COLLABORATIONS

Modern biology is demanding increasing collaboration and teamwork, as scientists mount efforts to tackle increasingly difficult problems. Unfortunately, our evaluation system has not kept up with this modern trend, and arguably is a deterrent to collaboration, because of fears of not getting sufficient credit for career advancement. The CV/biosketch, which rules the evaluation marketplace, is an information-poor, "winner-take-all" system. The precious first author slot is reserved for the young person who "drove" the experimental and presumably the intellectual aspects of the projects. The last author slot is reserved for the principal investigator whose lab provided the main environment and resources for the project. However, many collaborative research situations are not so black and white. To correct for this, the starred "shared authorship" (first, second, or last) as well as "author contributions" at the end of the scientific article have been introduced. In some cases, the senior author might relinquish the last author spot to a young investigator who played a critical role. These innovations to the CV have been helpful but not sufficiently dramatic to change attitudes about scientific collaborations and ease concerns about career advancement. Thus, rewarding teamwork of young scientists in our evaluation process remains a difficult problem.

One powerful tool that might be further exploited is the letter of evaluation, which provides the richest contextual information on a candidate and can describe elements of teamwork and collaboration. We need to better articulate the importance of letters of recommendation to our graduate students and postdocs and emphasize how they weigh heavily in career advancement decisions. Collaborating with other labs also provides opportunities for a young scientist to get to know a senior investigator (other than his or her PI) who can write a letter on his or her behalf. In addition, we might even consider incorporating letters of recommendation in the first-time R01 reviews of junior faculty. In addition to providing context for contributions in collaborations, a letter of recommendation

might enable study sections to better evaluate the promise of a new, young investigator and help them to base decisions on more information than preliminary results and the number of postdoc publications, factors that are contributing to the increasing age at which investigators are obtaining their first R01s (Matthews *et al.*, 2011).

COMMUNITY AND EDUCATION AS FACTORS FOR ADVANCEMENT

Papers and grant support are the gold standard for promotion in many research-focused academic institutions; achievements in education and community service tend to be much less valued. While scholarly achievement and grants sustain the core mission of research institutions, education and community service also are important and creative endeavors; they contribute immensely to the culture of an institution and the future of our profession. These efforts should be respected and deserve more than lip service during a review for academic promotion. Academic evaluation predicated too narrowly on papers and impact factors steers young scientists away from educational/community activities if these activities contribute only minimally to their overall evaluation. This sends the wrong message to young scientists, especially at a stage when many desire both to be altruistic and to advance their careers.

What steps could be taken to increase the importance of education/community service in evaluations? This is an issue that needs to be addressed primarily by individual institutions and laboratories. The value of education and community service should be incorporated into the messages and priorities that department chairs communicate to their faculty; it should also be inherent in the values that senior faculty convey to their postdocs and students. "Impactful" has become an adjective that precedes "paper," but it should be used more broadly to describe other important ways in which scientists contribute to society and our profession. An initially high-flying paper might be largely forgotten a year later, while important educational/community activities can have ripple effects that influence an institution decades later.

CONCLUSION

As stewards of our profession, academic scientists have a collective responsibility to consider how to disseminate knowledge through publications and how to advance graduate students to postdocs, postdocs to assistant professors, and assistant professors to tenure and beyond. These processes are not out of our hands, predetermined, or immutable. What do we value? What are we rewarding? Are current metrics working, are they changing with the times, and are we investing sufficient time and effort in the evaluation process?

Consensus answers to these questions are unlikely, and perfect solutions may not be possible. Reducing our work and values to simple numbers (e.g., h-index) is unlikely to be the answer, and I have argued for a more complete and holistic examination of overall laboratory productivity, mentoring of students and postdocs, and other activities during peer review. Such additional information requires little time for PIs to provide and peer reviewers to read and helps to define more clearly the ideals our profession is striving to achieve. Beyond establishing ideals, the evaluation system is behind the stage, pulling the strings of how the business of the scientific enterprise is being conducted; its effects are potent, since it deals most directly with the passions and ambitions of individuals. The future of the scientific workforce, the training of future scientists, the ways in which we publish new knowledge, and the scientific culture of labs and institutions all are influenced by the metrics and process of evaluation. "Evaluating how we evaluate" is therefore an important issue for broader discussion, and it is time for us to take ownership of the process.

REFERENCES

Alberts BM (1985). Limits to growth: in biology, small science is good science. Cell 41, 337–338.

Ball P (2007). Achievement index climbs the ranks. Nature 448, 737.

Johnston M (2009). Reclaiming responsibility for setting standards. Genetics 181, 355–356.

Matthews KR, Calhoun KM, Lo N, Ho V (2011). The aging of biomedical research in the United States. PLoS One 6, e29738.

Olson MV (1995). A time to sequence. Science 270, 394–396.

Raff M, Johnson A, Walter P (2008). Painful publishing. Science 321, 36.

Sauermann H, Roach M (2012). Science PhD career preferences: levels, changes, and advisor encouragement. PLoS One 7, e36307.

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