DAVID MOHER, FLORIAN NAUDET, IOANA A. CRISTEA, FRANK MIEDEMA, JOHN P. IOANNIDIS, AND STEVEN N. GOODMAN

New Principles for Assessing

S C I E N

How should academic institutions best assess science and scientists? Burgeoning interest in this question accompanies a growing recognition of significant problems in how scientific research is conducted and reported. The right questions are not being asked; the research is not appropriately planned and conducted; reproducibility is lacking; and when the research is completed, results remain unavailable, unpublished, or selectively reported.

Such problems are connected to the processes by which scientists are assessed to inform decisions about their hiring, promotion, and tenure. Building, writing, presenting, evaluating, prioritizing, and selecting curriculum vitae is a prolific and often time-consuming industry for grant applicants, faculty candidates, and assessment committees. Institutions need to make decisions in an environment of limited time and constrained budgets. Many current assessment efforts consider primarily what is easily determined, such as the number and amount of funded grants and the number and citations of published papers.

Even for readily measurable aspects of a scientist's performance, though, the criteria used for assessment and decisions vary across institutions and are not necessarily applied consistently, even within the same institution. Moreover, many institutions use metrics that are well known to be problematic. For example, a large literature documents the problems with journal impact factor (JIF) for appraising citation impact. That faculty hiring and advancement at top institutions requires papers published in journals with the highest JIF (*Nature*, *Science*, *Cell*, etc.) is more than just a myth circulating among postdoctoral students. The JIF is still a benchmark that most institutions use to assess faculty or even to determine monetary rewards. But emphasis on the JIF does not make sense when only 10%-20% of the papers published in a journal are responsible for 80%-90% of a journal's impact factor.

More important, other aspects of research impact and quality for which automated indices are not available are ignored. For example, faculty practices that make a university and its research more open and available through data sharing or education could feed into researcher assessments. Few assessments

of scientists focus on the use of good or bad research practices, nor do currently used measures say much about what researchers contribute to society—the ultimate goal of most applied research. In applied and life sciences, the reproducibility of methods and findings by others is only now starting to be systematically evaluated. Most of the findings indicate substantial concerns. A former dean of medicine at Harvard University, Jeffrey Flier, has indicated that reproducibility should be a consideration when assessing scientists' performance.

Using more appropriate incentives and rewards may help improve clinical and life sciences and their impact at all levels, including their societal value. A number of existing efforts demonstrate the growing awareness of the need for reform as well as the range of approaches and ideas on the table. Large group efforts, including the Leiden Manifesto for Research Metrics and the Declaration on Research Assessment (DORA), both developed at academic society meetings, and both international in focus, are but two examples. Individual or small-group proposals for assessing scientists include one from a group led by Madhu Mazumdar at Mt. Sinai Health System that emphasizes the importance of rewarding biostatisticians for their contributions to team science. Scientific journals are a perhaps-surprising third source of ideas. Although they traditionally have been focused on—even obsessed with—promoting their JIFs, more enlightened and progressive journals are beginning to acknowledge the metric's limitations and to consider other metrics. Finally, newer efforts to improve quantitative metrics are ongoing in the vast and rapidly expanding field of scientometrics, where it seems that every metric has its strengths and weaknesses, including the possibility for "gaming," or manipulation by the investigator.

We reviewed 22 key documents critiquing the current incentive system. We extracted how the authors perceived the problems of assessing science and scientists, the unintended consequences of maintaining the status quo for assessing scientists, and details of their proposed solutions. We then convened an expert panel workshop in Washington, DC, in January 2017 to consider such existing efforts to improve the evaluation of "life and clinical" research scientists, to see how a broad spectrum of stakeholders view the strengths and weaknesses of these efforts, and to discuss whether new ways of assessing scientists should be considered. We have previously published a full description of the data analysis and group deliberations ("Assessing biologists for hiring, promotion, and tenure," *PLOS Biology* 16, no. 3 [2018], from which this essay is adapted). Here we want to highlight the six general principles that emerged from our work, each with research and policy implications. We believe that these principles have application across the broad scientific research enterprise.

T I S T S

PRINCIPLE 1. Contributing to societal needs is an important goal of scholarship. Focusing on research that addresses societal needs and the impact of research requires a broader, outward view of scientific investigation. This principle is based on academic institutions in society, how they view scholarship in the twenty-first century, the importance of patients and the public, and social action. If promotion and tenure committees do not reward these behaviors or penalize practices that diminish the social benefit of research, maximal fulfillment of this goal is unlikely.

PRINCIPLE 2. Assessing scientists should be based on evidence and indicators that can incentivize best publication practices. Several new "responsible indicators for assessing

scientists" (RIAS) were proposed and discussed. These include whether scientists register a research project prior to its conduct (including registered reports) and how the researchers share methods and results (including code and materials) of research. The indicators also include the reproducibility of research; contributions to peer review; alternative metrics (e.g., uptake of research by social media and print media) assessed by several providers, such as Altmetric.com; and the impact of the research. Such indicators should be measured objectively and accurately, as publication and citation data are currently. Some assessment items, such as reference letters from colleagues and stakeholders affected by the research, cannot be converted into objective measurements, but they might still be used when formally investigating their value.

As with any new measures, RIAS characteristics need to be studied in terms of ease of collection, their frequencies and distributions in different fields and institutions, the kind of systems needed to implement them, and their usefulness in both evaluation and modifying researcher behaviors, and the extent to which each may be gamed. Different institutions could and should experiment with different sets of RIAS to assess their feasibility and utility.

Ultimately, if there were enough consensus around a core set of responsible indicators, institutional research funding could be tied to their collection, as is happening with the successful implementation of Athena SWAN (Scientific Women's Academic Network) for advancing gender equity, which has been highly successful in the United Kingdom.

One barrier to implementation of any RIAS model is its potential to affect current university rankings, such as the Times Higher Education World University Rankings. Productivity, measured in terms of publication output, is an important input into such rankings. Participants felt that any RIAS evaluation could be included in or used as an alternative to university ranking methods, which are themselves problematic.

PRINCIPLE 3. All research should be published completely and transparently, regardless of the results. Academic institutions could implement policies in the promotion process to review complete reporting of all research, to penalize noncompleted or nonpublished research, or both—particularly regarding clinical trials, which must be registered. For nonclinical research, there is a need to reward other types of openness, such as the sharing of datasets, materials, software, and methods used, and to provide explicit acknowledgment of their exploratory nature, when appropriate.

Finding fair ways to reward team endeavors is critical, given the growing collaborative nature of research, which bibliometrics cannot properly assess. For example, some promotion and tenure committees largely disregard work for which the faculty candidate is not the first or senior author. Conversely, citation metrics that do not correct for multiple coauthorship, and thus reward authors who are just appearing in long author mastheads, can result in inappropriately high citation metrics.

PRINCIPLE 4. Openness should be encouraged and assessed in terms of dissemination and use of research methods, data, and results by others. Researchers can share their data, code, procedures (methods and materials), and code in various ways, such as in open-access repositories and preprint servers. A growing number of journals and publishers are supporting this process by endorsing and implementing the transparency and openness promotion guidelines. Groups that rank universities can also support this principle by sharing the underlying data used to make their assessments.

PRINCIPLE 5. Additional investments are necessary in research to provide the necessary evidence to guide the development of new assessment criteria and to evaluate the merits of existing ones—that is, research on research. Funders are well positioned to make such investments, and some, particularly in Europe (e.g., the Netherlands Organization for Scientific Research, the Wellcome Trust) have already started.

PRINCIPLE 6. Researchers should be rewarded for intellectual risk-taking that might not be reflected in early successes, grants, or publications. The need for young researchers to obtain their own funding early often results in a conservatism that is inimical to groundbreaking work at a time when the scientists might be the most creative. Changing assessments to evaluate and reward such hypotheses might encourage truly creative research. It is also possible to conduct some forms of research with limited funding.

Moving forward

A challenge when introducing any of these principles, or other new ideas, is how best to operationalize them. One tool that could help, TrialsTracker, enables institutions from around the world (with more than 30 trials) to monitor their trial reporting. Although the tool has limitations, it has a low barrier to implementation and provides a useful and easy starting point for institutional audit and feedback. Promotion and tenure committees could receive such data as part of annual faculty assessment. They could also ask scientists to modify their CVs to incorporate information about where they have registered their research, to indicate whether they have participated in a journal's registered reports program, and to add a citation of the completed and published study. For each new initiative, it is important to generate evidence,

ideally from experimental studies, on whether it leads to better outcomes.

Good, rigorously conducted evaluation should not come at the expense of stifling creative "blue-sky" research primarily aimed at understanding biologic processes. Such efforts should also be rewarded in assessment processes.

Current systems sometimes reward scientific innovation, but if the goal is to improve research reproducibility, ways must be found to reward scientists who focus on it. A scientist who detects analytical errors in published science and works with the authors to help correct the errors needs to have such work recognized. This benefits the original scientists, study participants in the original research, the journal publishing the original research, the field, and society. The authors of the original report could include documentation, perhaps in the form of an impact letter, attesting to the value of the reproducibility efforts, which could be included in the evaluation portfolio.

High-quality practice guidelines are evidence-based, typically using systematic reviews as one of their foundational building blocks. Similar evidence-based approaches also will need to be developed for assessing scientists. Although it has attracted some criticism, the United Kingdom's Research Excellence Framework is a step in this direction. The metrics marketplace is large and confusing. Institutions can choose or pick metrics with an evidence base and endorsed by reputable organizations. For example, the US National Institutes of Health sponsored development of the RCR (Relative Citation Ratio), and Leiden University's Centre for Science and Technology Studies developed the SNIP (Source-Normalized Impact per Paper). Regardless of which approach is adopted, evidence on the accuracy, validity, and impact of indicators is necessary.

Even if best practices for appraising scientists can be identified, achieving widespread adoption will be a major challenge. Ultimately, this may depend on institutional values, which might be elicited from the institution's faculty. Junior faculty may put a high value on open-access publications. If open access were to become part of RIAS and included in faculty assessments, the institution would need to support open-access fees more broadly or find other ways to promote a culture of openness (e.g., use of preprint servers or institutional repositories). Committed support from leadership and senior faculty would be needed to implement policy. Finally, implementation for some of the six principles should be easier if stakeholders work collaboratively.

Institutional promotion and tenure committee guidelines are not easily available to outside researchers, although there is an effort under way to compile them. If institutions made them available, this information could be used as a baseline to gauge changes in criteria and also to disseminate institutional innovations. Institutions can also examine their own rewards and promotion practices to understand how

Even if best practices for appraising scientists can be identified, achieving widespread adoption will be a major challenge.

their high-level criteria are being operationalized and to see the effect of criteria such as counting the number of first and last author publications. Funders can also make widely available what criteria they use to assess grant applicants.

Whether implemented at the local or national level, changes in assessment criteria should be fully documented and made openly available. Institutions making changes to their promotion and tenure criteria and faculty assessment should implement an evaluation component as part of the process. Evaluations using experimental approaches are likely to provide the most internally valid results and may offer greater generalizability. Novel study designs such as the stepped wedge cluster or interrupted time series might be appropriate for assessing the effects of individual or multiple department promotion and tenure committees' uptake of new assessment criteria for scientists, together with audit and feedback. These data can inform the development of new and effective systems.

How the academic community evaluates scientists reflects what it values most—and least—in the scientific enterprise and powerfully influences scientists' behavior. Widening the scope of activities worthy of academic recognition and reward will likely be a slow and iterative process. The principles here could serve as a road map for change. Although the collective efforts of funders, journals, and regulators will be critical, individual institutions will ultimately have to be the crucibles of innovation, serving as models for others. Institutions that monitor what they do and the changes that result would be powerful influencers of the shape of the collective scientific future.

David Moher *is a clinical epidemiologist and associate professor in the clinical epidemiology program and the director of the Centre for Journalology at Ottawa Hospital Research Institute.* ***Florian Naudet*** *is an associate professor of therapeutics at the University of Rennes in France.* ***Ioana A. Cristea*** *is an associate professor at Babeş-Bolyai University in Romania.* ***Frank Miedema*** *is a professor of immunology and the dean and vice-chairman of the board at the University Medical Center Utrecht in the Netherlands.* ***John P. Ioannidis*** *is a professor at Stanford University and co-director of the Meta-Research Innovation Center at Stanford (METRICS).* ***Steven N. Goodman****, co-director of METRICS, is the associate dean of clinical and translational research and a professor at Stanford University.*