Can Scientific Quality Be Quantified?  
Joseph Loscalzo, MD, PhD

“… The first essential step in the direction of learning any subject is to find principles of numerical reckoning and practicable methods for measuring some quality connected with it. I often say that when you can measure what you are speaking about and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.”

—Baron William Thomson Kelvin

There was a time in the not too distant past when the value of a scientific observation was determined simply by peer response to its communication. Through local lectures, presentations at national and international scientific meetings, and, ultimately, publication, the scientific community became aware of a new finding and assessed its potential importance and impact. Put another way, the quality of a scientific observation was determined by the extent of acclamation of the community within which it took root: Good scientists knew good science when they saw it. Full stop.

Article see p 1038

With the advent of the digital revolution, the development of international databases of published articles, and the ready ability to analyze these data sets, things changed. Beginning in the 1950s, and cooperatively exploiting their natural tendency to measure phenomena in the world around them, members of the scientific and information communities began to conceive of bibliometric indices that might be used to define the impact of a scientific publication and the journal in which it was published. An initial impetus for developing the first of these parameters, the science citation index, was to assist librarians in managing bibliographic control and costs effectively. The citation index quantifies the number of citations a particular publication receives, and this information is used, in turn, to calculate a journal-specific parameter, the journal impact factor. The impact factor is defined as the average number of citations received per paper published in a specific journal during the preceding 2 years. These 2 parameters have since evolved differently from their original intention: Both are now used as quantifiable measures of quality, of the scientist and of the journal in which the scientist publishes.

Before the era of the personal computer, these bibliometric indices were not widely available to the scientific community. Only librarians and journal editors had access to the information on an annual basis (for a fee). With growth in personal computing and electronic access to these searchable parameters, their importance took on new life as they became widely accepted “objective” measures of quality. For example, the citation index can now readily be explored by anyone working in an organization that subscribes to Thomson Reuters Institute of Scientific Information Web of Knowledge, Science Citation Index Expanded. Using this Web-based tool (or others, including Google Scholar and Scopus), one can ascertain the total number of times an author’s work (individual publication or total publications) has been cited in other publications, the number of citations per year for a given publication, the total number of citations per year for a journal’s published articles, and the impact factor of the journals in which the articles have been published. As a department chair and journal editor, I have personally witnessed the evolution of these indices over the past decade into a type of scientific currency. For example, I find them increasingly incorporated in curricula vitae and letters of reference for promotion, even when not requested. In some countries, these indices are now being used formally as objective indicators of suitability for promotion or increased financial compensation.

Similarly, the impact factor has become the central currency in the journal realm. Authors wish to publish in journals with the highest impact factors, as doing so will improve their personal analytics and the professional opportunities afforded them as a result. Every journal editor works feverishly to improve his or her journal’s impact factor because it is viewed by publishers as an index of journal quality and success, determining the extent to which the journal is resourced by its sponsoring organization or publisher. Despite a widely held belief to the contrary, the impact factor bears little, if any, relationship to advertising revenue raised: Advertisers primarily use the journal’s individual subscription volume to make advertising venue decisions.

It is important to consider the values and limitations of the citation index and the impact factor as we consider the flaws in the logic of their broad application to the objective evaluation of scientists, scientific publications, and journals. The citation index is driven by the scientific community of which the publishing scientist is a part. If that community is comparatively large and active, comprising members who publish frequently, the likelihood that a published paper will be cited is greater than if that community is more limited in size and productivity. I term this limitation community productivity bias, a concrete example of which is the difference in the frequency of citations of large, multiauthored, industry-funded clinical trials in cardiovascular medicine or oncology compared with other clinical trials or basic studies.
In this issue of Circulation, McAlister and colleagues show that bibliometric indices not only undervalue basic and nonrandomized clinical studies (or overvalue large, randomized, clinical trials) but also fail to provide a clear view of the paths to critical cardiovascular discoveries. They also emphasize another important feature of high-quality biomedical research in particular: It should improve health. One can, therefore, argue that insofar as large, randomized, clinical trials definitively demonstrate the benefit of a therapy, the publication of those trials directly leads to improvements in health and well-being, and they are recognized as such with high citation frequency. The limitation of recognizing these definitive trials with higher citation indices than the preclinical research on which they are based, however, is that doing so fails to acknowledge adequately the fundamental advance promoted by that preclinical research. Counterarguments to this concern are that without the definitive clinical trial, the value of the original preclinical observation would not have been realized, and not infrequently, the clinical advance of a preclinical finding bears little, if any, relationship to that expected on the basis of preclinical work (eg, use of tumor necrosis factor antagonists in rheumatologic diseases rather than heart failure).

This more limited recognition of basic investigation compared with large clinical trials is highlighted by an analysis of the citation indices of the work for which Nobel Laureates received the Nobel Prize. In Garfield’s original analysis of citation frequency in 1967, he listed the 50 most highly cited authors at the time and pointed out that 2 of them subsequently received the Nobel Prize (M. Gell-Mann, ranked sixth, received the Nobel Prize in Physics in 1969; and D.H.R. Barton, ranked forty-first, received the Nobel Prize in Chemistry in the same year). By inference (but without formal statistical assessment), these data were inductively generalized to indicate that any scientist doing work recognized as sufficiently novel and ground-breaking is more likely to be highly cited than one who is not. This cursory analysis was interpreted as validation of the broad use of the citation index as a general measure of scientific quality. To explore the validity of this assumption in a bit more detail, I reviewed the work of 10 of the last 10 years’ Nobel laureates in physiology or medicine and ascertained the citation frequency of the seminal work for which each was granted the prize. To limit the halo effect of the prize itself, the publication of those trials directly leads to improvements in health and well-being, and they are recognized as such with high citation frequency. The limitation of recognizing these definitive trials with higher citation indices than the preclinical research on which they are based, however, is that doing so fails to acknowledge adequately the fundamental advance promoted by that preclinical research. Counterarguments to this concern are that without the definitive clinical trial, the value of the original preclinical observation would not have been realized, and not infrequently, the clinical advance of a preclinical finding bears little, if any, relationship to that expected on the basis of preclinical work (eg, use of tumor necrosis factor antagonists in rheumatologic diseases rather than heart failure).

This more limited recognition of basic investigation compared with large clinical trials is highlighted by an analysis of the citation indices of the work for which Nobel Laureates received the Nobel Prize. In Garfield’s original analysis of citation frequency in 1967, he listed the 50 most highly cited authors at the time and pointed out that 2 of them subsequently received the Nobel Prize (M. Gell-Mann, ranked sixth, received the Nobel Prize in Physics in 1969; and D.H.R. Barton, ranked forty-first, received the Nobel Prize in Chemistry in the same year). By inference (but without formal statistical assessment), these data were inductively generalized to indicate that any scientist doing work recognized as sufficiently novel and ground-breaking is more likely to be highly cited than one who is not. This cursory analysis was interpreted as validation of the broad use of the citation index as a general measure of scientific quality. To explore the validity of this assumption in a bit more detail, I reviewed the work of 10 of the last 10 years’ Nobel laureates in physiology or medicine and ascertained the citation frequency of the seminal work for which each was granted the prize. To limit the halo effect of the prize itself, I determined the average annual and cumulative citation frequencies from the time of publication to the year before the Nobel Prize was awarded. The results of this analysis showed that the average total citations for these 10 laureates’ seminal articles were 1382±1127 (mean±SD), ranging from 323 citations to 4332 citations. The average annual citation frequency was 59.7±57.7 (mean±SD), ranging from 6.1 to 176.0 citations per year. Importantly, the lower citation frequencies did not involve prizes given for more recent work, or for work confined to a smaller, less productive community of scientists. In addition, only 4 of these 10 Nobel laureates met criteria for the Institute of Scientific Information’s Highly Cited recognition. These data indicate that although many scientists of the highest caliber (as assessed by the Nobel Committee) publish articles that are highly cited owing to their novelty and impact, many do not. This group of scientists should serve as an adequate “positive control” for the predictive accuracy of these indices as surrogates for scientific quality. One can, therefore, conclude from this analysis, limited though it is, that these bibliometric indices have an unacceptably high false-negative rate for assessing scientific quality.

Many highly cited articles are reviews of topics of broad interest or descriptions of novel research methodologies with broad applicability to biomedical research (such as the Lowry protein determination, cited more than 300 000 times, and the Southern blot, cited more than 31 000 times). Although these highly cited methods have been used to advance the biomedical research enterprise broadly, to argue that their high citation frequency implies that their scientific value is akin to that of Nobel-quality work challenges credulity. I term this limitation the statistical obliviousness bias. To put these citation frequencies in quantitative perspective, and as emphasized by McAlister and colleagues, it is important to realize that from 1900 to 2005, only 0.5% of the ≈38 million published papers were cited more than 200 times, and half were never cited at all.4,6 It would be interesting to ascertain the distribution of the very highly cited papers among the categories of original research articles, large clinical trials, methods papers, and reviews to gauge the extent to which the conflation of these distinctions confounds the assumption of scientific quality implicit in the citation frequency.

A third limitation of the citation index is the problem of an author citing his own work, ie, the autocitation bias. One can imagine that under circumstances in which one’s suitability for promotion or increased compensation is dependent on citation frequency, there would be a tendency to cite one’s own work. The Thomson Reuters Institute of Scientific Information Web of Knowledge acknowledges this potential problem and allows citation frequency searches to be conducted with or without autocitations included.

A fourth limitation of the citation index is the degree to which it is influenced by the publication venue. This flaw is one in which limitations of the citation index and the impact factor overlap. Publications in journals with high impact factors are more likely to be cited than publications of similar observations in journals with lesser impact factors, as indicated by the analysis of Ioannidis on journal ecology. In addition, as the number of journals increases, cited articles tend to be more recent and published in fewer journals. These phenomena likely promote and sustain the differences among these journals’ impact factors. This tautological flaw I term the self-prophetic bias. A journal’s reputation as a venue of quality is perpetuated by the attraction and publication of papers of potentially high citation value to the scientific community; in other words, scientists believe a publication is of high value because the editor of journal X says that it must be for it to have been published in journal X. Notwithstanding the intimation of quality implicit in this self-prophesy, the correlation between journal impact factors and citation indices is, by some assessments, rather poor. The recently developed Eigenfactor score improves this correlation by weighting citations in journals with high impact factors more heavily than those in journals with low impact factors; however, as a result, it further promotes the self-prophetic bias by definition.

The last limitation of the citation index to which I will refer is that of indiscriminate equivalency among all coauthors of
multiauthored papers. For basic papers or small clinical studies in which all authors truly contribute importantly and equivalently, this practice is both reasonable and just. By contrast, in heavily multiauthored clinical trials or large genome-wide association studies, for example, coauthorship may simply reflect a modest (but probably important) cocontributor role; yet, equivalent citation weight is allotted these “minor” authors, a flaw I term the equivalent contributor bias. The challenges of assigning authorship credit are not insubstantial, as recently reviewed by Biagioli, whose careful analysis further weakens the assumption of author equivalence on which these bibliometric indices are based.

The impact factor itself is also riddled with many flaws that can be misleading to readers and to journal editors. From the editor’s perspective, if the goal is to improve the impact factor of a journal, one can decrease the overall number of published articles, decrease the number of original articles published, and increase the number of highly citable reviews published. There are countless examples of the success of this manipulative strategy (in fact, as I learned a few years ago in discussions with an associate editor of a competing journal, there are consultants who can be hired to improve a journal’s impact factor using these painfully obvious approaches). One can also add to these tactics encouraging an author to cite papers published in the journal in which they are to be published, a policy that borders on the unethical, in my view. These manipulative approaches are increasingly used by editors desperate to improve and sustain an impact factor of high value.

The critical and implicit assumption in the development of the impact factor as a measure of journal use is that published scientists are active in their field, their work undergoes rigorous peer and editorial review before publication, and therefore, by citing a peer’s publication, they serve as arbiters of scientific quality. The impact factor, then, is not so much a measure of what is being read but what should be read by the larger scientific community, including those scientists who publish frequently and those who do not. Assigning such influence to published scientists sustains their influence in their field, driving it in a particular direction that is (consciously or subconsciously) self-affirming at best (and self-serving in the extreme). The great majority of their scientific progeny carry on where they left off, moving the field forward incrementally, often without changing the underlying scientific paradigm on which it was built; only a very small minority of scientists advance a field by significantly shifting the paradigm on which it stands, as Kuhn originally pointed out. There are, of course, well-known examples of game-changing scientific observations that have come from disparate fields of investigative endeavor, but even these need to garner the support of an influential scientific champion from within the discipline for successful recognition. This flaw, then, reflects the geopolitics of science and its influence in directing or hindering scientific progress, an unfortunate human dimension to the lofty ideal of the purity of scientific objectivity.

Alternatives to the citation index have been developed, most notably the “h” (or Hirsch) index. This index is based on an individual’s distribution of citations among his or her published papers, and is defined as the number of papers with citation frequency \( \geq h \). The h index was proposed by Hirsch as a useful measure of an individual’s scientific output, reflecting both the number of publications and the number of citations per publication. The 10 Nobel laureates whose productivity I analyzed had an average h index of 62 \( \pm 24 \) (mean \( \pm \) SD), with a range of 22 to 93. In general, the h index will increase over time, with minimal impact of self-citation. Implicit in the h index, however, is the same assumption of quality that plagues the citation index, namely, that citation by peer scientists is a reasonable surrogate for quality. No alternative to the impact factor has yet gained traction, and as it becomes more widely institutionalized, both academically and commercially, it is unlikely that a substitute will be developed.

If Eugene Garfield were designing a metric for the first time in 2011 rather than in 1964, I am certain that he would use a parameter different from the impact factor. After all, an original intention was to give librarians a use statistic with which to make rational decisions in managing their collections. Clearly, the impact factor is a surrogate for use and, as recently shown, a not very good one at that. Use statistics can best be ascertained in the current era by counting e-page views, an approach widely used by advertisers in determining the e-venues for their advertisements. Darmoni and colleagues recently defined a “reading index,” which equals the ratio of e-page views of articles within a specific journal to overall e-page views. Important confounders of this index include whether or not the full-text version is available online and whether or not access is restricted. Using this ratio for 46 biomedical journals of widely variable impact factors in 1997, these authors found that there was no correlation between the impact factor for a journal and the reading index, which illustrates the inadequacy of the impact factor even for the purpose for which it was originally intended, ie, as a measure of the journal’s use and reader appeal. This observation was confirmed more recently in an analysis of open-access journal readership and citation frequency: In the first year after publication, more readers accessed the articles in open-access journals but did so without any commensurate increase in citation frequency.

On the basis of the evidence presented here, I assert that the citation index and the impact factor are only weakly correlated, at best, with true quality, were we to have an independent measure of quality against which to compare them. Defining the quality of a scientific contribution is not a trivial exercise. Historians and philosophers of science have opined on the ironic limitations of quantifying the contributions of a field that prides itself on objectively quantifying observations. Mazlish, for example, pointed out that “[s]cience’s own scientific nature … supports the effort to understand the quality of science in a Kelvin-like [ie, measurable] fashion …. The value of science indicators appears self-evident. The limits and limitations involved in the approach are less obvious … [and] can carry with [them] unintended harmful possibilities … [causing] us to overlook or undervalue other means of understanding … the quality of science.” What are these “other means of understanding” scientific quality to which we should turn? Can scientific quality truly be assessed objectively by practitioners of science? Or should we assume that scientists are blinded by the biases of the era in which they work, subject to the predilections founded on their own scientific experiences and...
beliefs? If one were to ask any scientific investigator for those attributes of a scientific observation that define it as high quality, each would likely respond similarly: The highest-quality science is enlightening; changes the direction of a field dramatically; and advances the field not incrementally, but in a major quantal fashion. Many leading prize committees, the Nobel Committee first among them, have been able to recognize such truly superior work; however, even these august deliberative bodies and the leading scientists who influence them occasionally err in their selections (eg, the 1950 Nobel Prize in Physiology or Medicine, in part, to Hench for his demonstration of the beneficial effect of corticosteroids in rheumatoid arthritis, later demonstrated to have now well-known adverse consequences when used long-term; and the 1984 Nobel Prize in Physiology or Medicine, in part, to Jerne for his idiotype-anti-idiotype network hypothesis for the generation of antibody specificity and suppression of autoimmunity, later precluded by clonal selection). These examples and others like them highlight the importance of hindsight in recognizing true quality in science. Much like for history itself, in which the passage of time is generally required to appreciate the importance and impact of an historical event, so, too, for science: Only with the passage of time can the importance of an observation be put in the proper unbiased context and its true value appreciated.

What, then, are we to do with these widely accepted bibliometric approaches to quantifying scientific quality? There are so many confounders that influence both the citation index and the impact factor that to rely on them exclusively as a measure of the quality of a scientist or of a scientific journal defies logic. Not only are there numerous flawed assumptions underlying them, but the statistics that are applied to them are used inappropriately, and their statistical validity has not yet been studied adequately to provide some measure of predictive accuracy. Yet, wedded as we are to measurement, and seeking a simple way to gauge a difficult property to parameterize, scientific quality, we are at the mercy of these poor substitutes for a true quality index. My urging is that we not take these measures too seriously, and that when we choose or are forced to do so, we take into consideration potential mitigating factors that reflect the flaws and biases described here. Owing to the extraordinarily rich publication databases that are now available and the ease with which they can be analyzed, these bibliometric indices should at the very least be modified to take into account some of the most obvious flaws and confounders. Modifications of the index could include normalizing for the size of the scientific community; differential weighting for original articles, reviews, methods papers, editorials, etc; and differential weighting for multiauthored papers based on the magnitude of the individual author’s contribution. Even with refinement, however, we are still likely to be left with a suboptimal measure of the scientific quality of an individual or of a journal if we persist in a quest for a single assessment parameter.

The best arbiters of the quality of science are its contemporary practitioners who choose to adopt new and potentially important findings heuristically into their own scientific rubric, the wisdom of whose choices is ultimately evaluated through the prism of historical hindsight. Good scientists know good science when they see it. And see it they must, because, as Joshua Lederberg stated, “[t]o flourish, science has many needs, but none … more vital than [the] responsible communication with history, society, and posterity embodied in what we casually call the scientific literature.”

Acknowledgments

The author wishes to thank Elliott Antman, Karen Barry, Jeremy Greene, Anita Loscalzo, Marc Pfeffer, and Joseph Vita for helpful comments and Susan Vignolo-Collazzo for expert technical assistance.

Disclosures

None.

References

21. Davis PM. Open access publishing, article downloads, and citations: randomized controlled trial. BMJ. 2008;337:a568.
Can Scientific Quality Be Quantified?
Joseph Loscalzo

Circulation. 2011;123:947-950
doi: 10.1161/CIRCULATIONAHA.111.020529
Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2011 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://circ.ahajournals.org/content/123/9/947