A New Scale for Rating Scientific Publications

Răzvan Valentin Florian[1]

[1] *florian@epistemio.com*
Epistemio, str. Saturn nr. 26, 400504 Cluj-Napoca (Romania); 20-22 Wenlock Road, London N1 7GU (UK); and
Romanian Institute of Science and Technology, str. Cireșilor nr. 29, 400487 Cluj-Napoca (Romania)

Introduction

Citation-based bibliometric indicators are increasingly being used for evaluating research. This reflects the need of decision-makers to increase the efficiency of allocating resources to research institutions and scientists, while also keeping manageable and cost-effective the evaluation process that grounds the allocation of resources. There often is much room of improvement in how bibliometric indicators are being used in practice. But even state-of-the art bibliometric indicators suffer of a fundamental problem when used for evaluating research: the citations they are based upon are influenced by many factors beyond the quality of cited publications (Bornmann & Daniel, 2008) and these indicators need to be tested and validated against what it is that they purport to measure and predict, which is expert evaluation by peers (Harnad, 2008).

A solution to this problem is aggregating online ratings provided post-publication by the scientists who read the rated papers anyhow, for the purpose of their own research. Online-aggregated ratings are now a major factor in the decisions taken by consumers when choosing hotels, restaurants, movies and many other types of services or products. It is paradoxical that in science, a field for which peer review is a cornerstone, rating publications on dedicated online platforms is not yet a common behavior. For example, if each scientist would provide one rating weekly, it can be estimated that 52% of publications would get 10 ratings or more (Florian, 2012). This would be a significant enhancement for the evaluative information needed by decision makers that allocate resources to scientists and by other users of scientific publications.

For collecting this kind of ratings, a rating scale should be defined. Here I present the choices made during the development of the scale used at Epistemio, an online platform for aggregating ratings and reviews of scientific publications (www.epistemio.com).

Purpose

The expected usage of these ratings is: first, in steering of science by decision-makers, i.e. choosing to whom to allocate resources (typically contributed publicly), such as institutional funding, grants, jobs, positions, tenure, among the institutions, scientists, fields of science, etc. that compete for them; and second, in helping scientists to prioritize and filter the publications that they choose to read or use. For the first purpose, it is important to be possible to aggregate ratings across the set of publications of an individual, of a group of scientists or of an institution; and to be able to use the individual or aggregated ratings to rank the assessed entities. This implies that ratings should be unidimensional. While publications may be assessed across a number of characteristics, such as quality of research, quality of presentation, novelty, and interest, collecting individual ratings across all these dimensions reduces the response rates, and it is not clear how these multidimensional ratings may be aggregated into a scalar one. Therefore, it is desirable that an overall rating that reflects the overall properties of a publication is collected independently of ratings regarding individual characteristics of the publication. Collecting the latter may be left optional. This paper focuses on the overall rating.

What should be rated, exactly?

When experts are asked to rate a publication, the property that should be rated must be named. What is exactly this property? A proper discussion of this issue should analyze the foundations of scientific research, being outside the scope of the present paper. A different way of posing the problem is starting with the needs of expected users of the ratings, which were mentioned above. Typical desired properties of publications (and, therefore, of the results presented in these publications) that are mentioned in the context of steering of science is quality, importance, relevance, and impact. For usability purposes, the text of the question to raters should be kept brief; therefore, a choice must be made among the various wordings that may be used. Importance, long-term societal and scientific relevance, and long-term societal and scholarly impact seem to have similar semantics. Quality seems to be a complementary property: a publication may present potentially important results, but methodology and/or presentation may lack quality, therefore raising uncertainties about the real value of the publication; and a publication may be of high quality while the potential importance is low. We have thus chosen to use the wording “scientific quality and importance” for defining the variable that the ratings are supposed to estimate.

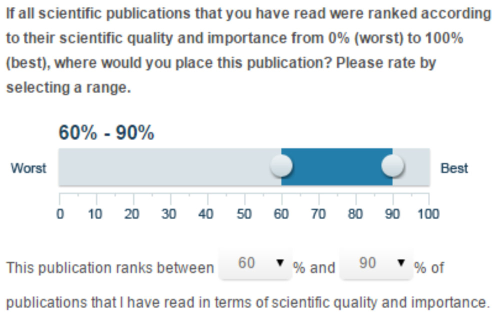
Scale type and range

Online ratings typically take the form of a five-star or ten-star discrete scale: this standard has been adopted by major players such as Amazon, Yelp, TripAdvisor and IMDb. However, these types of scales are likely not being able to measure well the quality and importance of scientific publications, because of the likely high skewness of the distribution of values of this target variable.

Let us consider the number of citations of scientific publications as a relevant proxy for the quality and importance of publications. About 44% of publications in Web of Science have zero citations, and the median number of citations is about 1, yet there is one paper having more than 305,000 citations and 148 papers having more than 10,000 citations (Van Noorden, Maher, & Nuzzo, 2014). In the case of patents, where the monetary value is defined by markets, the top 0.8% were valued at more than 1,000 times the median (Giuri et al., 2007). Let us assume that the main properties of these distributions generalize to the variable we want to measure, i.e. the maximum value can be of about 3 to 5 orders of magnitude larger than the median value. Therefore, a scale of 5, 10 or even 100 discrete categories cannot represent well this variability if the values that the scale represents vary linearly across categories. A logarithmic scale would be suitable, but it is psychologically difficult for most people to estimate values across so many orders of magnitude and to place them on a logarithmic scale.

A solution to this conundrum is asking experts to assess not the absolute value of the target variable, but its percentile rank. Then, the maximum value (100%) is represented by a number just 2 times larger than the median (50%), rather than several orders of magnitude larger. For usability and computational reasons, we limited the precision of the scale to 1%. Theoretically, this limits the capacity of indicating differences between top papers; in the case of the number of citations, in the top 1% the value varies from several hundreds to hundreds of thousands. In practice, test-retest reliability tends to decrease for scales with more than 10 response categories; users consider that a scale with 101 response categories allow them to best express their feelings adequately, but its ease and speed of use is slightly lower than of scales with 11 categories or less (Preston & Colman, 2000).

Because of the skewness of the distribution of absolute values, it is likely that experts are able to discriminate the percentile ranking of high quality papers better than the one of low quality papers. The confidence in rating papers also depends on how close the topic of the publication overlaps the expertise of the rater. For these reasons, raters should be able to express their uncertainty. Therefore, we allowed experts to give the rating as an interval of percentile rankings, rather than a single value. The rating is collected through a graphical interface representing the interval with sliding ends (Fig. 1). For ease of use on mobile devices, the interval can also be expressed using numerical selectors. A review may be associated to the rating, for explaining and supporting the rating.



If all scientific publications that you have read were ranked according to their scientific quality and importance from 0% (worst) to 100% (best), where would you place this publication? Please rate by selecting a range.

60% - 90%

This publication ranks between 60 ▾ % and 90 ▾ % of publications that I have read in terms of scientific quality and importance.

Figure 1. The Epistemio® rating scale for scientific publications.

Acknowledgments

This work was supported by a grant of the Romanian National Authority for Scientific Research, CNDI–UEFISCDI, project number PN-II-PT-PCCA-2011-3.2-0895.

References

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? *Journal of Documentation*, *64*(1), 45-80.

Florian, R. V. (2012). Aggregating post-publication peer reviews and ratings. *Frontiers in Computational Neuroscience*, 6(31).

Giuri, P., Mariani, M., Brusoni, S., Crespi, G., Francoz, D., Gambardella, A., et al. (2007). Inventors and invention processes in Europe: Results from the PatVal-EU survey. *Research Policy*, *36*(8), 1107–1127.

Harnad, S. (2008). Validating research performance metrics against peer rankings. *Ethics in Science and Environmental Politics*, *8*, 103–107.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(2000), 1-15.

Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature*, *514*(7524), 550–553.