Research evaluation for computer science

Bertrand Meyer (ETH Zurich)
Christine Choppy (LIPN, UMR CNRS7030, Université Paris 13)
Jørgen Staunstrup (IT University of Copenhagen)
Jan van Leeuwen (Utrecht University)

Academic culture is changing. The rest of the world, including university management, increasingly assesses scientists; we must demonstrate worth through indicators, often numeric. While the extent of the syndrome varies with countries and institutions, La Fontaine's words apply: "*not everyone will die, but everyone is hit*". Tempting as it may be to reject numerical evaluation, it will not go away. The problem for computer scientists is that assessment relies on often inappropriate and occasionally outlandish criteria. We should at least try to base it on metrics acceptable to the profession.

In discussions with computer scientists from around the world, this risk of deciding careers through distorted instruments comes out as a top concern. In the US it is mitigated by the influence of the Computing Research Association's 1999 "best practices" report[1]. In many other countries, computer scientists must repeatedly explain the specificity of their discipline to colleagues from other areas, for example in hiring and promotion committees. Even in the US, the CRA report, which predates widespread use of citation databases and indexes, is no longer sufficient.

Informatics Europe (http://www.informatics-europe.org), the association of European CS departments, has undertaken a study of the issue, of which this article is a preliminary result, whose views commit the authors only. For ease of use the conclusions are summarized through ten concrete recommendations.

Our focus is evaluation of individuals rather than departments or laboratories. The process often involves many criteria, whose importance varies with institutions: grants, number of PhDs and where they went, community recognition such as keynotes at prestigious conferences, best paper and other awards, editorial board memberships. We mostly consider a criterion that always plays an important role: *publications*.

Research evaluation

Research is a competitive endeavor. Researchers are accustomed to constant assessment: any work submitted — even, sometimes, invited — is peer-reviewed; rejection is frequent, even for senior scientists. Once published, a researcher's work will be regularly assessed against that of others. Researchers themselves referee papers for publication, participate in promotion committees, evaluate proposals for funding agencies, answer

[1] For this and other references, and the source of the data behind the results, see a fuller version of this article at http://se.ethz.ch/~meyer/publications/cacm/research_evaluation.pdf.

institutions' requests for evaluation letters. The research management edifice relies on assessment of researchers by researchers.

Criteria must be fair (to the extent possible for an activity circumscribed by the frailty of human judgment); openly specified; accepted by the target scientific community. While other disciplines often participate in evaluations, it is not acceptable to impose criteria from one discipline on another.

Computer science

Computer science concerns itself with the representation and processing of information using algorithmic techniques. (In Europe the more common term is *Informatics*, covering a slightly broader scope.) CS research includes two main flavors, not mutually exclusive: Theory, developing models of computations, programs, languages; Systems, building software artifacts and assessing their properties. In addition, domain-specific research addresses specifics of information and computing for particular application areas.

CS research combines aspects of engineering and natural sciences (in Systems) as well as mathematics (Theory and Systems).This diversity is part of the discipline's attraction, but also complicates evaluation.

Across these variants, CS research exhibits distinctive characteristics, captured by seminal concepts: algorithm, computability, complexity, specification/implementation duality, recursion, fixpoint, scale, function/data duality, static/dynamic duality, modeling, interaction… Not all scientists from other disciplines realize the existence of this corpus. Computer scientists are responsible for enforcing its role as basis for evaluation:

> 1. Computer science is an original discipline combining science and engineering. Researcher evaluation must be adapted to its specificity.

The CS publication culture

In the Computer Science publication culture, prestigious conferences are a favorite tool for presenting original research — unlike disciplines where the prestige goes to journals and conferences are for raw initial results. Acceptance rates at selective CS conferences hover between 10 and 20%; in 2007-2008:

- ICSE (software engineering): 13%.
- OOPSLA (object technology): 19%.
- POPL (programming languages): 18%.

Journals have their role, often to publish deeper versions of papers already presented at conferences. While many researchers use this opportunity, others have a successful career

based largely on conference papers. It is important not to use journals as the only yardstick for computer scientists.

Books, which some disciplines do not consider important scientific contributions, can be a primary vehicle in CS. Asked to name the most influential publication ever, many computer scientists will cite Knuth's *The Art of Computer Programming*. Seminal concepts such as Design Patterns first became known through books.

2. A distinctive feature of CS publication is the importance of selective conferences and books. Journals do not necessarily carry more prestige.

Publications are not the only scientific contributions. Sometimes the best way to demonstrate value is through software or other artifacts. The Google success story involves a fixpoint algorithm: Page Rank, which determines the popularity of a Web page from the number of links to it. Before Google was commercial it was research, whose outcome included a paper on Page Rank and the Google site. The site had — beyond its future commercial value — a *research* value that the paper could not convey: demonstrating scalability. Had the authors continued as researchers and come up for evaluation, the software would have been as significant as the paper.

Assessing such contributions is delicate: a million downloads do not prove scientific value. Publication, with its peer review, provides more easily decodable evaluation grids. In assessing CS and especially Systems research, however, publications do not suffice:

3. To assess impact, artifacts such as software can be as important as publications.

Another issue is assessing individual contributions to multi-author work. Disciplines have different practices (2007-2008):

- *Nature* over a year: maximum coauthors per article 22, average 7.3.
- *American Mathematical Monthly*: 6, 2.
- OOSPLA and POPL: 7, 2.7.

Disciplines where many coauthors are the norm use elaborate name ordering conventions to reflect individual contributions. No such culture exists in CS:

4. The order in which a CS publication lists authors is generally not significant. In the absence of specific indications, it should not serve as a factor in researcher evaluation.

Bibliometry

In assessment discussions, numbers typically beat no numbers; hence the temptation to reduce evaluations to such factors as publication counts, measuring output, and citation counts, measuring impact (and derived measures such as indexes, discussed next).

While numeric criteria trigger strong reactions[2], alternatives have problems too: peer review is strongly dependent on evaluators' choice and availability (the most competent are often the busiest) and does not scale up. The solution is in combining techniques, subject to human interpretation:

> 5. Numerical measurements such as citation counts must never be used as the sole evaluation instrument. They must be filtered through human interpretation, particularly to avoid errors, and complemented by peer review and assessment of outputs other than publications.

Measures should not address volume but impact. Publication counts only assess activity. Giving them any other value encourages "write-only" journals, speakers-only conferences, and Stakhanovist research profiles favoring quantity over quality.

> 6. Publication counts are not adequate indicators of research value. They measure productivity, but neither impact nor quality.

Citation counts assess impact. They rely on databases such as ISI, CiteSeer, ACM Digital Library, Google Scholar. They, too, have limitations:

- Focus. Publication quality is just one aspect of research quality, impact one aspect of publication quality, citations one aspect of impact.
- Identity. Misspellings and mangling of authors' names lose citations. If your name is Krötenfänger, do not expect your publications to be counted right.
- Distortions. Article introductions heavily cite surveys. The milestone article that introduced NP-completeness has far fewer citations than a later tutorial.
- Misinterpretation. Citation may imply criticism rather than appreciation. Many program verification articles cite a famous protocol paper — to show that their tools catch an equally famous error in the protocol.
- Time. Citation counts favor older contributions.
- Size. Citation counts are absolute; impact is relative to each community's size.
- Networking. Authors form Mutual Citation Societies.
- Bias. Some authors hope (unethically) to maximize chances of acceptance by citing PC members.

[2] David Parnas: *Stop the Numbers Game — Counting papers slows the rate of scientific progress*, in *Comm. of the ACM*, vol. 50, no. 11, November 2007, pages 19-21, available at http://tinyurl.com/2z652a. Parnas mostly discusses counting publications, but deals briefly with citation counts.

The last two examples illustrate the occasionally perverse effects of assessment techniques on research work itself.

The most serious problem is data quality; no process can be better than its data. Transparency is essential, as well as error reporting mechanisms and prompt response (as with ACM and DBLP):

> 7. Any evaluation criterion, especially quantitative, must be based on clear, published criteria.

This remains wishful thinking for major databases. The methods by which Google Scholar and ISI select documents and citations are not published or subject to debate.

Publication patterns vary across disciplines, reinforcing the comment that we should not judge one by the rules of another:

> 8. Numerical indicators must not serve for comparisons across disciplines.

This rule also applies to the issue (not otherwise addressed here) of evaluating laboratories or departments rather than individuals.

CS coverage in major databases

An issue of concern to computer scientists is the tendency to use databases that do not adequately cover CS, such as Thomson Scientific's ISI Web of Science.

The principal problem is what ISI counts. Many CS conferences and most books are not listed; conversely, some publications are included indiscriminately. The results make computer scientists cringe. Niklaus Wirth, Turing Award winner, appears for minor papers from indexed publications, not his seminal 1970 Pascal report. Knuth's milestone book series, with an astounding 15,000 citations in Google Scholar, does not figure. Neither do Knuth's three articles most frequently cited according to Google.

Evidence of ISI's shortcomings for CS is "internal coverage": the percentage of citations of a publication in the same database. ISI's internal coverage, over 80% for physics or chemistry, is only 38% for CS.

Another example is Springer's *Lecture Notes in Computer Science*, which ISI classified until 2006 as a journal. A great resource, LNCS provides fast publication of conference proceedings and reports. Many are excellent, some not. Lumping all into a single "journal" category was absurd, especially since ISI omits top non-LNCS conferences:

- The International Conference on Software Engineering (ICSE), the top conference in a field that has its own ISI category, is not indexed.

- An LNCS-published workshop at ICSE, where authors would typically try out ideas *not yet ready* for ICSE submission, was indexed.

ISI indexes *SIGPLAN Notices*, an *unrefereed* publication devoting ordinary issues to notes and letters and special issues to proceedings of such conferences as POPL. POPL papers appear in ISI — on the same footing as a reader's note in a regular issue.

The database has little understanding of CS. Its 50 most cited CS references include "*Chemometrics in food science*", from a "*Chemometrics and Intelligent Laboratory Systems*" journal. Many CS entries are not recognizable as milestone contributions. The cruelest comparison is with CiteSeer, whose Most Cited list includes many publications familiar to all computer scientists; it has *not a single entry in common* with the ISI list.

ISI's "highly cited researchers" list includes many prestigious computer scientists but leaves out such iconic names as Wirth, Parnas, Knuth and all the ten 2000-2006 Turing Award winners but one.

Since ISI's proprietary, closed process provides no clear role for community assessment, the situation is unlikely to improve.

The inevitable deficiencies of alternatives pale in consideration:

> 9. In assessing publications and citations, ISI Web of Science is inadequate for most of CS and must not be used. Alternatives include Google Scholar, CiteSeer and (potentially) ACM's Digital Library.

Anyone in charge of assessment should know that attempts to use ISI for CS will cause massive opposition and may lead to outright rejection of *any* numerical criteria, including more reasonable ones.

Assessment formulae

A recent trend is to rely on numerical measures of impact, derived from citation databases, especially the **h-index**, the highest n such that $C(n) \geq n$, where $C(n)$ is the citation count of the author's n-th ranked publication. Variants exist:

- The **individual h-index** divides the h-index by the number of authors, better reflecting individual contributions.

- The **g-index**, highest n such that the top n publications received (together) at least n^2 citations, corrects another h-index deficiency: not recognizing extremely influential publications. (If your second most cited work has 100 citations, the h-index does not care whether the first has 101 or 15000.)

The "Publish or Perish" site[3] computes these indexes from Google Scholar data.

Such indexes cannot be more credible than the underlying databases; results should always be checked manually for context and possible distortions.

It would be as counter-productive to reject these techniques as to use them blindly to get definitive researcher assessments. There is no substitute for a careful process involving complementary sources such as peer review.

Assessing assessment

Scientists are taught rigor: submit any hypothesis to scrutiny, any experiment to duplication, any theorem to independent proof. They naturally assume that processes affecting their careers will be subjected to similar standards. Just as they do not expect, in arguing with a PhD student, to impose a scientifically flawed view on the sole basis of seniority, so will they not let management impose a flawed evaluation mechanism on the sole basis of authority:

10. Assessment criteria must themselves undergo assessment and revision.

Openness and self-improvement are the price to pay to ensure a successful process, endorsed by the community.

This observation is representative of our more general conclusion. Negative reactions to new assessment techniques deserve consideration. They are not rejections of assessment per se but calls for a professional, rational approach. The bad news is that there is no easy formula; no tool will deliver a magic number defining the measure of a researcher. The good news is that we have ever more instruments at our disposal, which taken together can help form a truthful picture of CS research effectiveness. Their use should undergo the same scrutiny that we apply to our work as scientists.

[3] www.harzing.com/resources.htm#/pop.htm.