

Evaluation of Faculty at IST – a case study

1. The Context

These days we all hear about the use and misuse of numbers and indicators in the evaluation and ranking of mathematical research. Although I have always paid some attention to what individuals and associations in the Mathematics community have been reporting on these issues, this seemed to be a distant reality until recently, when I have come to experience face to face with it. My colleagues at the Department must have felt the same, when a new evaluation procedure of the faculty, largely based on “numerology”, came into place.

Instituto Superior Técnico (IST) is the top science and engineering school in Portugal and its Mathematics Department is a large research department, with a faculty of around 100 active researchers covering many areas of mathematics, including both pure and applied mathematics. IST is a public school, which has to obey by the legislation and regulations set up by the Ministry of Science and Education. In 2009 a new legislation for public universities went into place, which for the first time imposed a regular evaluation of faculty. Evaluations will occur every 3 years, and the results of each individual evaluation are used to determined salary increases and the teaching load for the following 3-year period. The new legislation also opened the possibility of evaluating faculty in the period 2004-2009, during which salary increases had been frozen.

The need for a proper evaluation of faculty was long recognized by all active researchers in Portugal. Before the new regulations, salary increases were automatic for everyone after each 3-year period of activity (although salary increases had been frozen by the government since 2003, when the country first experience some economic troubles) and the teaching load was the same for every faculty member (but it varied from school to school). The new regulations set up some general rules and guidelines for evaluation of the faculty (e.g., 4 levels of performance: Poor, Good, Very Good, Excellent) as well as its consequences (e.g., a teaching weekly load between 6 and 9 hours), but left most of the details of the evaluation method to be determined by each university and school.

At IST the Scientific Council elaborates most regulations that apply to the faculty. The entire faculty at IST elects the Council, so its membership reflects the size and the number of its departments. At the time the rules for evaluation were set up, among the 25 members of the Council there were only two members from the Mathematics Department, a computer scientist and myself. The Council, after an intense debate, which lasted for around 1 year and included public consultation, approved the new rules for evaluation of the faculty. These new

rules had a large consensus among most of the Engineering Departments, faced some objections from the Physics Department, and had some strong objections from a part of the Mathematics Department, including myself (but not the other representative of the department in the Council).

In the end, the Council approved an evaluation system to be described below, which is largely based on numbers/indicators, with a small input of evaluation by peers. Since IST was the first school to implement this procedure, most science and engineering schools in universities throughout the country adopted this evaluation system, or slight variations of it. I don't know of any other country where an institution has adopted a similar evaluation procedure, but I suspect this may happen in the near future. In this article we argue that such evaluation methods are not effective. We use the very same indicators that these methods seek to measure to invalidate them. It is also a warning about a science fiction movie that can come to a theater near you.

2. The Evaluation System at IST

In order to understand how much in this evaluation system is based on indicators and how much depends on evaluation by peers, we will have to get a bit into the details of the IST evaluation system.

Faculty at IST is evaluated along 4 different aspects:

1. Teaching (which includes lecturing, pedagogical publications, advising of students, etc.)
2. Research (which includes research publications, participation and leadership of research projects, etc.)
3. Transfer of knowledge (which includes outreach activities, organization of conferences, patents, etc.)
4. Administration (which includes participation in school committees and councils, performance in departmental and school jobs, etc.)

The evaluation system gives an individual a certain number of points in each of these four different aspects, as it will be describe below. Then, depending on the level of the position, they are combined with certain weights, so for example, a full professor has weights:

- Teaching: 20%-40%
- Research: 40%-60%
- Transfer of knowledge: 5%-30%
- Administration: 10%-20%

These are ranges, not fixed values. Since faculty members can have different profiles (some are more research oriented, others are more teaching oriented, etc.), the weighted sum is subject to an optimization to maximize the final score, so that the sum of the weights is 1.

In the end, each faculty member gets a total score which, depending on the range where the score falls, corresponds to one of 4 level of performances: Poor (0-20), Good (20-50), Very Good (50-100) and Excellent (100 or larger). The total score

remains confidential, and only the level of performance is public and relevant for salary increases and other issues (e.g., determining future teaching load).

It remains to explain how each of the 4 different topics is evaluated. This is where a combination of indexes and peer-to-peer evaluation comes in. We will concentrate here on the topic "Research". The other 3 topics are evaluated on a similar fashion, but have certain peculiarities, related to the country's university system, which would require some longer explanations.

Evaluation of Research is divided into two different criteria:

- a) Research Publications;
- b) Research Projects;

In each of these criteria the score is a product of two factors:

$$S = Q \times M \quad (1)$$

where:

- Q is a **qualitative factor**, which by default is 1. For each faculty member it is nominated an evaluator (a peer of the same scientific area of equal or higher rank), which can attribute a different value to the qualitative factor, among the values $Q=0.5, 0.75, 1.0, 1.25, 1.5$;
- M is a **quantitative factor**, computed from a formula.

For example, for Research Publications the formula for the quantitative factor is as follows:

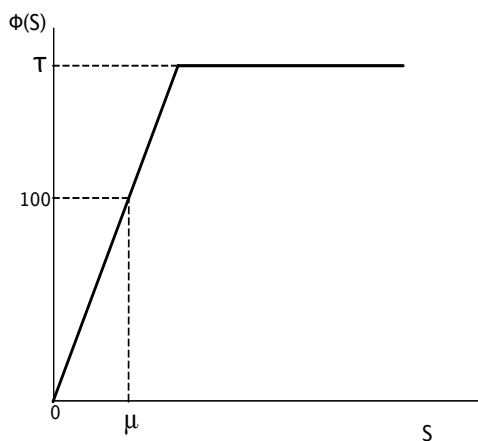
$$M = \sum_{i=1}^N A_i \times \left(T_i + \frac{C_i}{\rho} \right) \quad (2)$$

where:

- N is the number of publications that have appeared during the 3-year evaluation period;
- A_i is a factor that accounts for the number of authors of the i -publication ($3/2$ for one author, $5/4$ for two authors, $3/n$ for $n \geq 3$ authors).
- T_i is the type of the i -publication; so $T_i = 5.5$ for a research monograph, $T_i = 3, 1.75$ or 0.3 for a paper published in a journal of type A, type B or type C (for this purpose journals were classified into 3 different types), etc.
- C_i is the number of citations in ISI Web of Knowledge (Thomson Reuters) of the i -publication, excluding self-citations.
- ρ is factor that takes into account the reference number of citations of a given area, but which was taken equal to 5 for all areas in the first evaluation period.

The formula for the quantitative factor M for the criteria Research Projects has a similar nature and takes into account the number of projects, the value of the projects, the number of members of the project, the type of the project, whether is was a project coordinator or not, etc.

Since the values of the score S for different criteria (either in the same topic or in



different topics) are not commensurable, a transfer function $\Phi(S)$ is applied. The transfer function is a piecewise linear function that depends on two parameters, the target μ and the ceiling τ , as depicted in Figure 1. The specific values of these parameters vary with the criteria. So, for example, for “Research Publications” the values were set as $\mu = 4.5$ and $\tau = 600$. In principle, the values of these parameters could vary with the scientific area, but for the evaluation period 2004-2009 they were taken all to be the same.

Figure 1 - Transfer function

After the rescaling obtained by applying the transfer function, the various criteria are combined with fixed weights. So the final score in “Research” is a combination of 75% research publications with 25% research projects.

There are two important aspects in the evaluation system used at IST, which deserve to be analyzed. One is that the model allows for parameters that can be adapted to different fields of science, and even to different areas in each field. The other aspect is that it allows for limited peer-to-peer evaluation. These two aspects will be discussed in the next section.

3. Evaluation of different fields of research

IST has, besides the math department, a physics department and seven engineering departments. The members of each department are grouped in scientific areas. For example, the Mathematics Department has 8 areas: Algebra and Topology, Real and Functional Analysis, Differential Equations and Dynamical Systems, Geometry, Mathematical Physics, Numerical Analysis and Applied Analysis, Probability and Statistics, Logic and Computation. Comparing researchers in different areas is a very challenging exercise, to say the least. For example, we all know that different areas of Mathematics have different publications traditions. Even in a given area, there are researchers with different profiles, so any evaluation system that attempts to value the number of published papers is most likely to be a failure. Still it is possible to give some evidence for how much these differences can vary between different disciplines.

Soon after the new evaluation system at IST went into place I became head of the math department. I had expressed my objections of the evaluation system, while I was a member of the scientific council, and I was now very worried about the implementation of the new system. As head of the department I proposed different values for some of the parameters (e.g., the for the value of the target

μ for the criterion of research publications) but the central administration decided to keep the same values for all the school for the evaluation period 2004-2009. As a consequence, when the results came out, the mathematics department had 60% of its faculty evaluated with “Excellent”, compared with an average of 75% for the whole school. The top departments were the chemistry and biological engineering department (95% with “Excellent”) and the physics department (89% with “Excellent”). As a response to this, I created a working group in the department to promote an international “benchmark” of the relevant parameters in the model among different areas of science and engineering.

The first problem that the working group faced was the lack of data at the European level. Detailed statistics about science and engineering in Europe seem to be lost somewhere in between the national science councils and the mess of European directorates that fund various aspects of science. Answers to simple questions like: how many PhDs in a given area (say mathematics, chemistry, physics, etc.) are there working in academia in Europe, or how much funding does Europe devote to research in mathematics or in physics, don’t seem easy to answer. The last complete detailed report on science in Europe that we could find dates back to 2003, in the times of Commissioner Philippe Busquin¹.

By contrast, in the USA the National Science Foundation publishes every other year the report “Science & Engineering Indicators” and makes them freely available at <http://www.nsf.gov/statistics/>, including all tables and data in a format that can readily be used. Using these reports the working group easily produced evidence for how much indicators can vary in different fields of science. First, data was collected on the decade 1997-2006, for the various fields of science, for the academic sector (since the focus was in the evaluation of an academic institution, the study excluded the industry and the private sector). The working group considered the following aspects:

- Scientific productivity;
- Impact;
- Funding;

The next paragraphs describe some of the findings of the working group².

3.1 Productivity by field

In order to compare productivity in different fields of science, one can determine the ratio between the number of articles and the number of researchers in a given field. Associating a researcher or an article to a given field maybe problematic. To avoid this problem the degree field was defined as the field of the researcher, while data about articles were taken from the *Science Citation Index (SCI)* and the *Social Sciences Citation Index (SSCI)*, which assign fields to articles.

¹ “Third European Report on Science & Technology Indicators”, European Communities, 2003.

² The full report (in Portuguese) is available at www.math.ist.utl.pt/~rfern/BenchmarkReport.pdf.

A first indicator is given by the ratio between the number of articles and the number of researchers in the field:

$$\frac{\text{number of articles in the field}}{\text{number of researchers in the field}} \bigg/ \frac{\text{number of articles in mathematics}}{\text{number of researchers in mathematics}}$$

Table 1 presents this ratio normalized to mathematics.

Field	1997	1999	2001	2003	2006	Average
Mathematics	1.0	1.0	1.0	1.0	1.0	1.0
Physical Sciences	5.7	4.7	4.3	4.9	5.0	4.9
Computer sciences	2.2	1.7	1.7	1.7	1.6	1.8
Engineering	1.6	1.5	1.4	1.6	1.7	1.6
Life sciences	6.9	5.4	4.9	5.3	5.0	5.5
Psychology	1.6	1.3	1.1	1.3	1.3	1.3
Social sciences	1.1	0.9	0.8	0.9	0.9	0.9

Table 1 - Ratio articles/researchers in the USA, by field, gauged to mathematics³

Note that the data in Table 1 does not represent the average number of articles that a researcher in a given field publishes but rather the average number of articles per capita in a given field, gauged to mathematics.

In order to estimate the average number of articles that a researcher in a given field publishes one observes that:

$$\text{average number of articles per researcher} = \frac{\text{number of articles}}{\text{number of researchers}} \times (\text{average number of authors per article})$$

Therefore, in order to determine the average number of articles that a researcher in a given field publishes we need data about the average number of authors per article in a given field. This is presented in Table 2.

Field	1997	1999	2001	2003	2006	Average
Mathematics	1.8	1.8	1.9	1.9	2.0	1.9
Physical Sciences	3.6	3.8	4.0	4.2	4.6	4.0
Computer sciences	2.2	2.4	2.5	2.6	2.8	2.5
Engineering	3.0	3.2	3.3	3.4	3.6	3.3
Life sciences	3.5	3.7	3.9	4.1	4.4	3.9
Psychology	2.4	2.6	2.7	2.8	3.0	2.7
Social sciences	1.6	1.6	1.7	1.8	1.9	1.7

Table 2 - USA authors per S&E articles, by field ⁴

³ Source: Science and Engineering Indicators 2010 - Appendix table 5-15 (full-time faculty with S&E doctorates employed in academia by degree field) and Appendix table 5-42 (S&E articles from academic sector by field).

⁴ Source: Science and Engineering Indicators 2010 - Table 5-16 – Authors per S&E articles, by field. For the fields “Physical Sciences” and “Engineering” it was taken the average of their subfields and the missing years were obtained by linear interpolation.

Finally, Table 3 shows the average number of articles of a researcher in a given field, gauged to mathematics.

Field	1997	1999	2001	2003	2006	Average
Mathematics	1.0	1.0	1.0	1.0	1.0	1.0
Physical Sciences	11.6	9.7	9.1	10.8	11.8	10.6
Computer sciences	2.8	2.2	2.2	2.3	2.3	2.3
Engineering	2.8	2.6	2.5	2.9	3.2	2.8
Life sciences	13.8	11.0	10.3	11.3	11.2	11.5
Psychology	2.3	1.8	1.6	1.9	2.0	1.9
Social sciences	1.0	0.8	0.8	0.9	0.9	0.8

Table 3 - Average number of articles of a USA researcher, by field, gauged to mathematics

This data seem to suggests that in the USA academic institutions a computer scientist publishes on average twice more articles than a mathematician, an engineer publishes 3 times more articles than a mathematician, and a physicist a chemist or a biologist publishes 11 times more articles than a mathematician.

The fact that these numbers vary so much across different fields, together with the fact that boundaries between fields are usually diffuse, leads to the conclusion that even different areas in the same field should also have very different publications profile. Of course, collecting data like the one above for different areas in the same filed is almost impossible. Needless to say, we all know in our own areas of research that even top researchers have different publications profile. This should be enough to prevent using number of papers for evaluation purposes but, unfortunately, as we saw before this is not the case.

3.2 Impact factors by field

Other parameters that enter into formula (2) can be similarly examined. For example, to see how impact factors can vary across fields, we consider the ratio:

$$\frac{\text{number of citations in the field}}{\text{number of articles in the field}} \bigg/ \frac{\text{number of citations in Mathematics}}{\text{number of articles in Mathematics}}$$

The values of this ratio for different fields are presented in Table 4.

Field	1995	1997	1999	2001	2003	Average
Mathematics	1.0	1.0	1.0	1.0	1.0	1.0
Physical sciences	4.4	5.0	4.6	4.1	3.7	4.3
Engineering and Computer Science	1.7	1.8	1.6	1.6	1.4	1.6
Life sciences	6.3	7.8	6.9	6.6	6.0	6.7
Social sciences and Psychology	4.4	5.6	5.1	5.2	4.8	5.0

Table 4 - Citations per article by field, gauged to mathematics ⁵

⁵ Source: Science and Engineering Indicators 2006 - Appendix table 5-24: Worldwide citations of U.S. scientific articles, by field.

This data suggests that, on average, an article in engineering receives 1.6 times as much citations than one in mathematics, an article in Physics or in Chemistry receives 4 times as much citations than one in mathematics, and an article in the life sciences receives 7 times as much citations than one in mathematics.

It is also interesting to compare other indexes related to citations, which exhibit the different nature the fields, and which also has strong consequences when it comes to evaluation. It is common to use citations as a measure of impact, and this is usually limited to some period of time. However, depending on the field, articles make take different time to receive citations. For example, in ISI Web of Knowledge one finds the following indices for journals:

- *Immediacy Index*: the number of citations to an article in the journal during the year it is published.
- *Cited Half-Life*: the average time that takes an article in the journal to receive half of its citations.
- *Citing Half-Life*: the median age of the articles cited by the articles published in the journal.

Table 5 shows the aggregate values of these indices for the journals in each field.

Field	Aggregate Immediacy Index	Aggregate Cited Half-Life	Aggregate Citing Half-Life
Mathematics	0.160	>10	>10
Physical sciences			
Astronomy	1.461	6.6	7.1
Chemistry	0.543	6.4	8.1
Physics	0.553	7.1	8.1
Computer sciences	0.298	8.0	7.0
Engineering			
Bioengineering/biomedical	0.410	5.9	7.3
Chemical	0.306	6.9	8.3
Civil	0.290	7.0	8.4
Electrical	0.195	7.2	7.0
Mechanical	0.184	7.7	9.6
Life sciences			
Agricultural sciences	0.232	8.2	9.2
Biological sciences	0.731	6.7	7.2
Psychology	0.448	>10	9.0
Social sciences			
Economics	0.246	>10	9.0
Political science	0.193	9.2	8.5
Sociology	0.158	>10	9.7

Table 5 – Aggregate indices of impact in time for journals in a given field⁶

This data clearly shows that articles in mathematics take much longer to be cited and have a far longer influence than articles in most other fields of science. Needless to say, an article can stay unnoticed and become influent years after its

⁶ Source: Thomson Reuters, Science Citation Index and Social Sciences Citation Index.

publication. Inside each field, the number of researchers in a given area can vary widely. In mathematic there are clear differences between pure and more applied areas. All this, invalidates any formula that attempts to measure individual research by using citation data.

3.3 Funding by field

Administrators tend to value the amount of funding a researcher is able to raise. In IST evaluation system, there is a formula for research projects where the amount of funding of each project is parte of the data. Is it equally easy or difficult to researchers in different areas to raise funds? We all suspect that applied areas have a more generous funding than basic areas of research. After all, applied areas require expensive equipment and labs, which can justify large differences in funding.

Unfortunately, as have already mentioned above, there is not enough data available about Europe that allows for a clear picture of funding of different fields of science⁷. On the other hand, expenditures of research and development of different sectors in USA, by field, is readily available in the NSF Science and Engineering Indicators Report. Similarly to what we did above for articles and citations, we can estimate the funding per researcher in a given area, relative to mathematics, by computing the ratio:

$$\frac{\text{expenditures in R\&D in the field}}{\text{number of researchers in the field}} \bigg/ \frac{\text{expenditures in R\&D in mathematics}}{\text{number of researchers in mathematics}}$$

Table 6 below shows the values of the ratio for different fields.

Field	1997	1999	2001	2003	2006	Average
Mathematical sciences	1.0	1.0	1.0	1.0	1.0	1.0
Physical sciences	6.9	6.6	5.9	6.4	6.1	6.4
Computer sciences	11.1	10.9	10.1	10.3	7.7	10.0
Engineering	8.4	8.4	8.0	8.9	8.4	8.4
Life sciences	11.6	11.2	11.1	12.4	11.7	11.6
Psychology	0.9	0.9	0.9	1.2	1.0	1.0
Social sciences	1.4	1.3	1.3	1.4	1.1	1.3

Table 6 - Expenditures in R&D per researcher, by field, gauged to mathematics ⁸

The data in this table suggests that, on average, in the USA academic institutions, a physicist, an engineer or a biologist receives, respectively, 6 times, 8 times and 11 times more funding than a mathematician.

In Europe, the current trend is to establish research programs and fund research projects with visible applications to the real world. This makes life hard for

⁷ The EMS executive committee is currently leading an effort to produce data of funding of mathematics in Europe.

⁸ Source: Science and Engineering Indicators 2010 - Appendix table 5-6: Expenditures for academic R&D, by field

researchers working in more fundamental research. Evaluating research output of a person by using a formula measuring raised funds is highly unfair.

3.4 Peer-to-peer evaluation

The IST evaluation system allows for a limited peer-to-peer evaluation (see (1)). The proponents of this evaluation system readily recognized that even a sophisticated formula by itself couldn't possibly measure the quality of research of a person. The IST evaluation system limits the possible intervention of the evaluator because the authors of the system were afraid that the evaluator could simply overwrite the output of the formulas and invert what the "numbers say".

The data presented above shows not only that the use of formulas to measure quality of research is inadequate, but also that the order of magnitude of the error is such that peer-to-peer evaluation is needed and cannot be limited in any way. One may argue that the order of magnitude in the differences in the tables above is due to the fact that we treat different fields. However, even inside the same field (indeed, even in the same area of research) the differences can be huge. A Field medalist can have few publications compared to a largely unknown author.

The problems with evaluation by formulas tend to worsen after the persons under evaluation know how the formulas are built: they will act to potentially increase the numbers. The classical example is to try to multiply the number of papers by publishing shorter papers. This increase in the numbers does not necessarily mean an increase in the quality of research. Actually, it is the reason for many frauds with articles and journals. Again, the only way to fight this is through peer-to-peer evaluation.

Certainly, peer-to-peer evaluation is a human procedure and so it has many flaws. The way to minimize them is through public scrutiny, making available any relevant data and reports used in an evaluation. Just like democracy, peer-to-peer evaluation is not a perfect system but we don't know of any other system that is more perfect.

4. Concluding Thoughts

Although inside mathematics there is a large consensus about the dangers of using numbers and indicators in the evaluation and ranking of research, this is not so outside mathematics. Since the practices and the cultural environment in other sciences vary widely from mathematics, it may be helpful to have at hand data of the type collected here, to convince our colleagues in other departments why it is dangerous to do so. At IST this seems to have produced some results: the administration of the school received the document produced by the math department's working group and is considering modifying the evaluation systems accordingly.