STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

The Diversity of European Research Evaluation Systems[1]

Michael Ochsner[*], Emanuel Kulczycki[**] and Aldis Gedutis[***]

[*] *ochsner@gess.ethz.ch*
DGESS, ETH Zurich, Mühlegasse 21, Zürich, 8001 (Switzerland) and
FORS, University of Lausanne, Géopolis, Lausanne, 1015 (Switzerland)

[**] *emek@amu.edu.pl*
Adam Mickiewicz University, Scholarly Communication Research Group, Szamarzewskiego 89c, 60-568 Poznań (Poland)

[***] *aldis.gedutis@ku.lt*
Centre for Studies of Social Change, Klaipėda University, Minijos 153, 93185, Klaipėda (Lithuania

Introduction

Universities have undergone profound changes in the last decades. A shift towards more accountability and to "new public management" practices in the administration of universities took place and led to an increase of the share of project funds in some countries and to the introduction of performance-based funding systems (PRFSs) in others (see, e.g., Hicks, 2012; Lepori, Reale & Spinello, 2018). In all countries, research evaluation's importance increases. However, while research evaluation is centralized in some countries, evaluation is organized at the institutional level only in others. Thus, the importance of research evaluation and how it is organised varies across countries. Several typologies have been suggested to get an overview of research evaluation systems (Coryn et al., 2007; Geuna & Martin, 2001;2003; Hicks, 2010; 2012; Lepori, Reale & Spinello, 2018; von Tunzelmann & Mbula, 2003). Nonetheless, they all have some weaknesses. First, they only cover a restricted amount of countries, usually those for which information is publicly available in English. Second, they do not reflect whether such systems allow for adaptations of evaluation methods on the discipline level, i.e. no use of metrics in the social sciences and humanities (SSH); third, they often focus on mostly on financial impacts of the evaluation or focus exclusively on performance-based funding systems.

In this paper, we present a typology of national research evaluation systems in Europe, Israel and South Africa that sheds light on the complex issue of national differences in the organisation of research evaluation.

Data and Method

We use the data of a two-round Delphi survey among specialists in research evaluation as a basis of our analysis. The data was gathered in the context of the COST-Action 15137 "European Network for Research Evaluation in the Social Sciences and Humanities (ENRESSH)". The design of the analysis consists of five steps. First, a preliminary set of dimensions to classify research evaluation systems based on the existing typologies (Coryn et

[1] This article is based upon work from COST Action ENRESSH (CA15137), supported by COST (European Cooperation in Science and Technology) http://www.cost.eu/

al., 2007; Geuna & Martin, 2001;2003; Hicks, 2010; 2012; von Tunzelmann & Mbula, 2003) was developed and expanded by additional dimensions by members of the Steering Group of the Action. In a second step, a survey based on these dimensions was developed and fielded among the 60 Management Committee members of the Action before the kick-off meeting in March/April 2016. This first round of the Delphi survey aimed at finding out whether the expanded dimensions are formulated in a meaningful way and whether additional aspects should be added to characterize the broad range of countries included in the study (see Galleron, Ochsner, Spaapen & Williams, 2017). In a third step, the results from the survey were used to adapt the formulation and selection of dimensions and aspects of research evaluation systems and to develop the questionnaire for the second Delphi round. The second round of the survey was fielded among all members of the COST-Action that grew to 132 members from 38 countries between May 2017 and July 2017.

The surveys were fielded among all members of the COST-Action ENRESSH, who are all specialists in research evaluation. The first survey was fielded just before the start of the Action, the second in the second year. We aimed at multiple answers from the countries for several reasons: first, the research evaluation systems are not clearly defined and it was the aim of our surveys to find better adapted dimensions and aspects for such systems. Therefore, we were interested in whether representatives of the countries agreed on the single dimensions. Second, research evaluation systems are very complex and difficult to understand (Hicks, 2012, Lepori et al, 2018); having an opinion of more than one person helps in gauging the results.

For the classification of national evaluation systems, we used Multiple Correspondence Analysis (MCA, see, e.g., Greenacre, 2007) as implemented in Stata 14.2, using Burt matrices as input. From the results, we plot the countries and variables in a two-dimensional map. We then construct types of national research evaluation systems from the map. These types are of course not homogeneous as each country has its own way of evaluating research. Rather, the types should be understood as "ideal types" in the Weberian sense (Weber 1904/1949), i.e. types are formed by certain characteristics of the phenomena of interest but are not corresponding to *all* characteristics, thus they are not real but abstract representations of the phenomena. Ideal types serve to map, systematize and simplify complex phenomena. Real representations, in our case evaluation systems, can then be classified and described using the characteristics of the ideal types.

Survey Response and Variables

The first round of the survey reached a high response rate: 43 respondents from 25 countries filled in the questionnaire, which corresponds to 72% on the individual and 79% on the country level. For ten countries, more than one answer is available. The main result was that the existing dimensions do not reflect all necessary dimensions and aspects of research evaluation systems: First, there was much disagreement within countries on the same dimension pointing to the fact that the dimension needs adjustment; second, the open comment fields were extensively used. Clearly, this is at least partly due to a more heterogeneous selection of countries than in the previous studies (for a more comprehensive analysis of the first survey round, see Galleron et al., 2017). Besides changes of formulation and additions of aspects to dimensions, the main change of the questionnaire was a split into three main topics, consisting of similar dimensions (if applicable): institutional evaluation, career promotion and grant evaluation. Even though the survey was explicitly only on the first topic, the comments made it clear that in many cases, respondents answered taking into account that there is also an important impact of the national career promotion system or they considered also grant evaluation. The expansion to three topics

led to a significant increase of the length of the questionnaire but also increased the clarity for the respondents.

The second round of the survey was answered by 72 respondents from 33 countries, which corresponds to a high response rate of 55% on the individual and 87% on the country level. For 17 countries, more than one answer is available. The results show that the dimensions and aspects were clearer. However, we had to exclude Belgium from our analysis as the different regions of Belgium differ significantly regarding evaluation. Nevertheless, the Belgian experts tried to answer for Belgium as a whole, which led to non-classifiable results. This leaves us with 68 respondents from 32 countries. The countries in our analysis are: Austria, Bosnia Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Hungary, Iceland, Ireland, Israel, Italy, Latvia, Lithuania, Macedonia, Malta, Montenegro, the Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, South Africa, Spain and Switzerland. Due to space restrictions, we do not list the set of dimensions and aspects of the questionnaire but limit the description to the variables used for the classification. For the analysis, we used dummy-coded variables from different sets of variables from the questionnaire. The variables cover a) the existence of a comprehensive national publication database, b) whether evaluation is linked to funding (PRFS component), c) whether metrics take an important part in evaluation d) whether there is an SSH discipline-specific evaluation, e) whether there is a push to English publications, f) whether gender issues are addressed in evaluations (i.e. maternity/paternity leave, longer time periods to achieve standards for parents, etc.), g) the existence of a national career promotion institution or procedure and h) whether there are specific grant programs dedicated to the SSH.

We used the following decision rules for the dummy coding: For each respondent, the occurrence of the aspect was coded as 1 if the answer included the aspect (e.g. if the respondent checked the answer "Both performance-based funding and formative evaluation are implemented but separate from each other", the variable funding was coded to 1). The country was attributed a 1 if the majority of the respondents from a country scored a 1 on the variable.

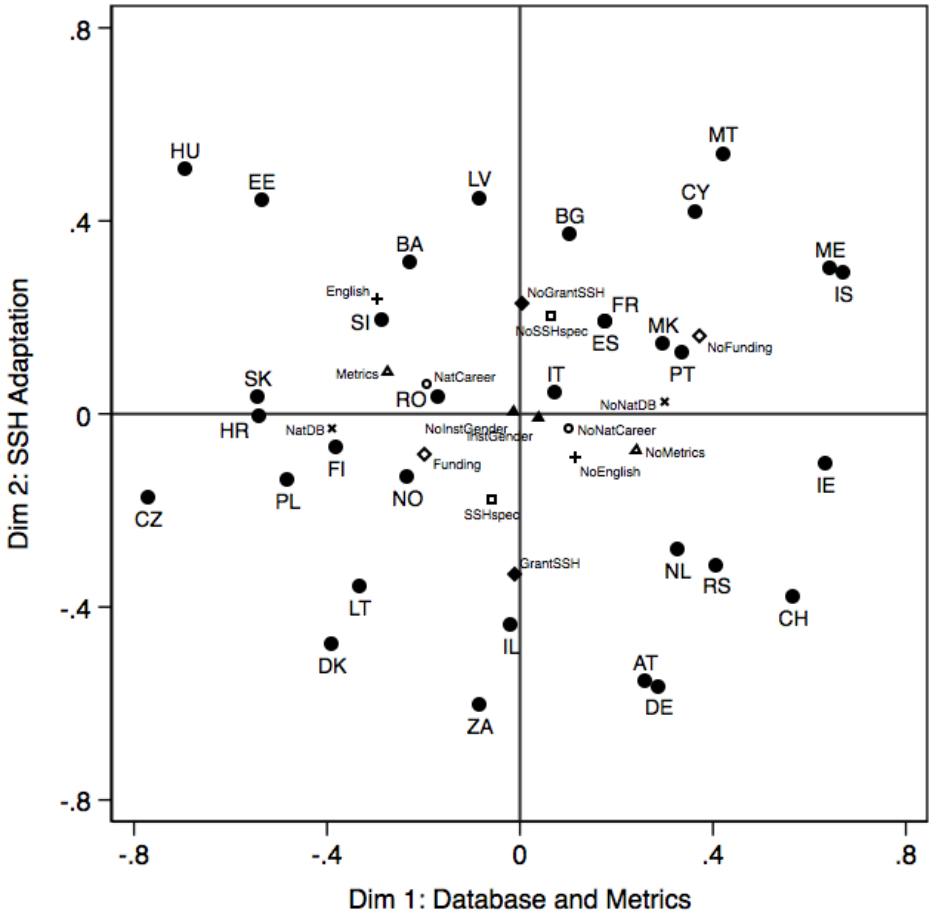
Classification

Figure 1 shows the map of the Multiple Correspondence Analysis. The two dimensions explain 34% of the total inertia, which is acceptable given the high number of variables and countries[2]. For the interpretation of the map, we first focus on the position of the variables (other symbols than full circles). While we are more interested in the space as such, the dimensions nevertheless reveal interesting information: The first dimension represents the metric component of evaluation: whether a national publication database exists, whether the principal method for evaluation uses metrics, whether funding is attached to evaluation. The second dimension is related to whether a system allows for adaptations to SSH research practices. Note that the more an item is placed towards the middle of the graph (around the origin), the closer it is to the means of the variables. Thus, such items do not add to the definition of the dimensions. It is notable that both ends of the variable Gender/NoGender are situated at the origin. This means that the dimensions do not really differ regarding the reflection of gender issues. Gender defines

[2] Multiple Correspondence Analyses underestimate the explained inertia systematically because they calculate the explained inertia from the whole matrix while only the explained inertia from the off-diagonal is of interest. Greenacre (2007) thus suggests using another, iterative algorithm, the so-called Joint Correspondence Analysis (JCA) that explains total variance more precisely but comes with other problems (e.g. the inertia explained does only make sense for all dimensions combined but not for single dimensions). As we are not interested in the amount of the explained inertia but rather in the placement of countries and variables in two-dimensional space to create a typology, we use the Burt matrix method. A JCA with the same data would yield a solution that explains 59% of total inertia. The visual representations do not differ substantially.

the third dimension of the MCA, adding another 10% of explained inertia. Due to restrictions of space, we do not include this third dimension in our typology as it consists only of one variable and the graphical display becomes less readable. However, it is a very notable result that the inclusion of gender issues in evaluation systems are not linked to the other characteristics of the evaluation system but builds the third dimension.

Figure 1. Map of Multiple Correspondence Analysis of national research evaluation systems.



Notes. Full circles represent countries, all other symbols represent dummy variables of characteristics of research evaluation systems. English/NoEnglish: system incentivises (or not) English language publications; (No)Funding: evaluation results affect funding; (No)GrantSSH: SSH-specific grant programmes; (No)InstGender: evaluation procedures reflect gender issues; (No)Metrics: main method of evaluation are metrics; (No)NatCareer: national career promotion procedure; (No)NatDB: national publication database; (No)SSHspec: SSH-specific institutional evaluation procedures.

Changing the focus to the placement of the countries in this two-dimensional space, we suggest 5 types of research evaluation systems, mainly linked to the quadrants but separating two types in the lower left quadrant. The first type, "non-metric, non-SSH" stands for national evaluation systems that do not have a national publication database, are not based on metrics, are not linked to funding and do not have SSH-specific procedures. Countries in that type include Cyprus (CY), France (FR), Iceland (IS), Macedonia (MK), Malta (MT), Montenegro (ME), Portugal (PT) and Spain (ES). The most representative country of this type is Iceland. All other countries deviate on one or two variables (the Southern European countries, for example, link funding to evaluation results). The second type, "non-metric, SSH-specific" consists of evaluation systems that do not have a national database, do not use metrics as a primary evaluation method, do not incentivise publications in English and have dedicated funding programs for SSH research. Countries in that type are Austria (AT), Germany (DE), Ireland (IE), the Netherlands (NL),

Serbia (RS) and Switzerland (CH). The prototype of such an evaluation system is Switzerland. The third type, "funding, non-metric" consists of evaluation systems using a national publication database, linking funding to evaluation results but the primary method of evaluation is peer review and the evaluation procedures are SSH-specific. Countries associated with this type are Lithuania (LT), Norway (NO) and South Africa (ZA). The main representative of this type is Norway. The fourth type "funding, metric" is characterized by using a national publication database, using metrics as a primary method for evaluation and linking evaluation results to funding while allowing for SSH-specific evaluation procedures and not incentivising publications in English. Countries in this type include Croatia (HR), Czech Republic (CZ), Denmark (DK), Finland (FI) and Poland (PL). Denmark represents this type best. Finally, the fifth cluster, "metric, English", stands for evaluation systems that have a national database in place, use metrics as a primary method of evaluation, link funding to evaluation results, do not allow for SSH-specific adaptations and incentivise English publications. Countries associated with this type include Bosnia Herzegovina (BA), Estonia (EE), Hungary (HU), Slovenia (SI) and Slovakia (SK). Estonia best represents this type.

Discussion and conclusions

Systematic research evaluation has become more and more important at universities. Some countries developed centralized national evaluation systems (see, e.g., Hicks, 2012), other countries actively refrained from centralizing and standardizing research evaluation but leave evaluation to institutions to best support their specific missions (e.g., Hasgall, Lanarès, Marion & Bregy, 2018), still other countries did centralize only some aspects of evaluation (see., e.g., Lepori et al., 2018). This led to a diverse landscape of research evaluation in Europe and beyond. Our analysis of research evaluation in 32 mostly European countries reveals that, indeed, countries have built quite unique evaluation systems. Nevertheless, some aspects can be identified that allow to classify evaluation systems.

In this paper, we suggest a typology that goes beyond existing classifications of research systems in that it a) includes a much broader range of countries, including countries, for which not much information is available to the English-speaking research community; b) takes into account that during the last years also the SSH are more concerned with evaluations but the commonly applied evaluation instruments do not fit their research practices; c) focuses not primarily on financial aspects.

Our data bases on 68 experts' assessments of the evaluation system in their own country. For the majority of the countries, the assessment bases on more than one expert. The results show that research evaluation systems are complex (see also Lepori et al., 2018). Experts do not always agree on all dimensions. This has several reasons but an important one is that implementations and practical applications are not always congruent with the formal definition. Another relevant reason is that evaluation systems consist of many components with different characteristics and that different experts might weigh the components differently. In this sense, this typology represents the evaluation experts' perceptions of the evaluation system in their own country.

Our empirical analysis suggests five ideal types of research evaluation: "non-metric, non-SSH" (with Iceland as the best representor), "non-metric, SSH-specific" (Switzerland, "funding, non-metric" (Norway), "funding, metric" (Denmark) and "metric, English" (Estonia). We also identified a third dimension orthogonal to those types, that reflects inclusion of gender issues in evaluations. Some countries do not fit into one type but are mixes of different types (Italy, Israel, Latvia, Romania).

The main result of this analysis is that the national organization of research evaluation system is a complex issue and the research evaluation landscape in Europe is diverse. Yet some components can be identified that define main types of research evaluation. A secondary result is that different types of research evaluation are linked to different conditions in countries. It is notable that the Southern European countries, the German speaking countries and the Nordic countries cluster together. This suggests that there is a link between research evaluation systems and historical or political structures. It is also important to note that some research-intensive and high performing countries (e.g., Germany, the Netherlands or Switzerland) follow a less metric but more adaptive approach while other countries try to increase their position in rankings using a metric approach that favours English publications (e.g. Hungary, Estonia, Bosnia Herzegovina). Thus, it can be concluded that evaluation systems should be adapted to the specific research situation in a country. Different evaluation systems create specific incentives and thus cause different effects or results. We therefore recommend, that designers of evaluation systems make a conscious link between the goals to achieve, the incentives to promote and the design of the evaluation system, rather than to strive to the unification of evaluation systems.

References

Coryn, C. L. S., Hattie, J. A., Scriven, M., & Hartmann, D. J. (2007). Models and Mechanisms for Evaluating Government-Funded Research: An International Comparison. *American Journal of Evaluation, 28*(4), 437–457. doi:10.1177/1098214007308290.

Galleron, I., Ochsner, M., Spaapen, J., & Williams, G. (2017). Valorizing SSH research: Towards a new approach to evaluate SSH research' value for society. *Fteval Journal for Research and Technology Policy Evaluation, 44*, 35–41. doi:10.22163/fteval.2017.274

Geuna, A., & Martin, B. R., (2001). University Research Evaluation and Funding: An International Comparison. *SPRU Electronic Working Paper Series* (Vol. 71).

Geuna, A., & Martin, B. R., (2003). University Research Evaluation and Funding: An International Comparison. *Minerva, 41*(4), 277–304. doi:10.1023/B:MINE.0000005155.70870.bd.

Greenacre, M. (2006). Correspondence Analysis and related methods in practice. In J. Blasius & M. Greenacre (ed.) *Multiple Correspondence Analysis and Related Methods* (pp. 3–40). London: Chapman & Hall.

Greenacre, M. (2007). *Correspondence Analysis in Practice* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Hasgall, A., Lanarès, J., Marion, A., & Bregy, J. (2018). *The programme 'Research performance in the humanities and social sciences'*. Report. Bern: swissuniversities.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy, 41*(2), 251–261. doi:10.1016/j.respol.2011.09.007.

Lepori, B., Reale, E., & Spinello, A. O. (2018). Conceptualizing and measuring performance orientation of research funding systems. *Research Evaluation*, advance access. doi:10.1093/reseval/rvy007

von Tunzelmann, N., & Mbula, E. K. (2003). *Changes in research assessment practices in other countries since 1999: final report*. Retrieved from http://www.ra-review.ac.uk/reports/prac/changingpractices.pdf (last access: 2018 04 13)

Weber, M. (1904/1949). Objectivity in Social Science and Social Policy, in E. A. Shils and H. A. Finch (ed. and trans.), *The Methodology of the Social Sciences* (pp. 49 - 112). New York, NY: Free Press.