Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca
anvur
National Agency for the Evaluation of Universities and Research Institutes

On the use of Bibliometric information for assessing articles quality: an analysis based on the third Italian research evaluation exercise

Marco Alfò[*], Sergio Benedetto[**], Marco Malgarini[***] , and Scipione Sarlo[***]

*Sapienza Università di Roma, **Politecnico di Torino, and ***ANVUR

STI 2017
Paris, September 8, 2017

Aims and scope

- Peer review is usually considered the main method for assessing the quality of research outputs
- The use of bibliometric indicators measuring the scientific impact can be a useful support
- Bibliometrics is a proxy measure of the concept of quality that can be fully assessed only by the expert judgement of *peers*
- Using only *peer review* in large research evaluation exercises may become very costly and almost unfeasible
- The degree according to which *bibliometrics* is a good proxy for *peer review*, and the problem concerning which indicator or combination of indicators should be used, are greatly disputed issues

Aims and scope

- Our aim is to shed some light on the matter, using a sample of articles, drawn from the third Italian research evaluation exercise (VQR), evaluated by both *peer review* and *bibliometrics*
- We will describe the Italian evaluation exercise, and afterward present the dataset that will be used in the analysis
- The relationship among peer review assessments and various bibliometric indicators is then thoroughly investigated by regression models
- Our conclusion is that the best proxy among those available for peer evaluation seems to be obtained by combining information from citations and journal impact

The Third Italian Research Evaluation Exercise (VQR 2011-2014)

- The Third Italian Research Evaluation Exercise (VQR) has been launched in June, 2015, and covers the period 2011-2014
- The exercise was in charge of the Italian National Agency for the evaluation of Universities and Research Institutes (ANVUR) - *ENQA Affiliates from September 2013*
- The VQR involved more than 60,400 researchers working in 96 Universities, 12 Public Research Organizations (PRO) and 27 other research bodies (participating on a voluntary basis)
- Over 118,000 publications have been submitted by researchers’ Institutions and evaluated by 16 Groups of Experts for Evaluation (GEV), involving 436 top-notch scientists

VQR 2011-2014:
Groups of Experts for Evaluation

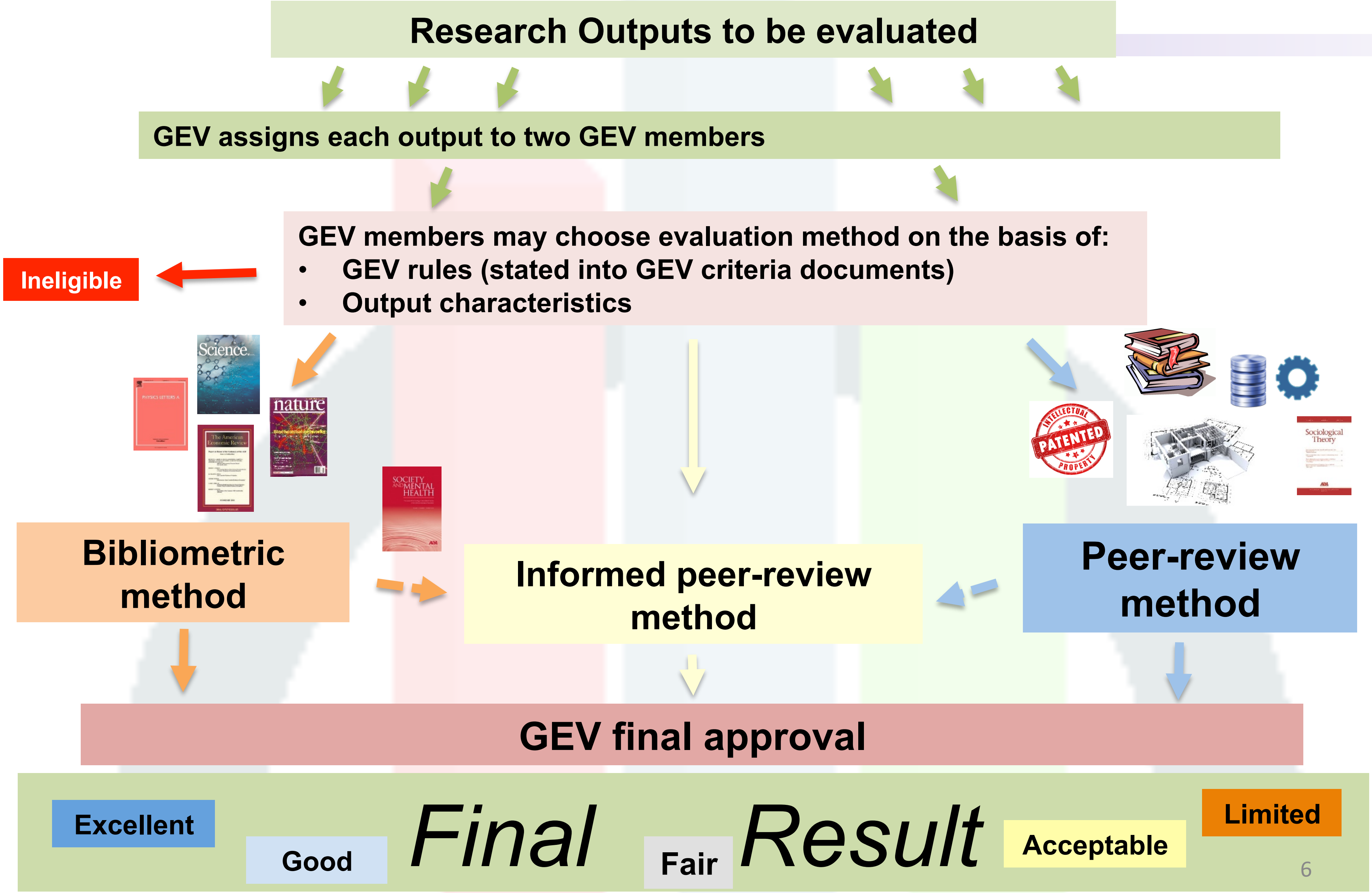
- GEV can be distinguished into two categories:
  - “Bibliometric GEV” (*mainly STEM areas*):

    1- Mathematics and Computer Sciences
    2 - Physics
    3 - Chemistry
    4 - Earth Sciences
    5 - Biology
    6 - Medicine
    7 - Agricultural and veterinary sciences
    8b - Civil Engineering
    9 - Industrial and Information Engineering
    11b – Psychology

  - “Non Bibliometric GEV” (*mainly social and humanities*):

    8a - Architecture
    10 - Ancient History, Philology, Literature and Art History
    11a - History, Philosophy, Pedagogy
    12 - Law
    13 - Economics and Statistics
    14 - Political and Social Sciences

VQR 2011-2014: the research outputs evaluation methods



Research Outputs to be evaluated

GEV assigns each output to two GEV members

GEV members may choose evaluation method on the basis of:

- GEV rules (stated into GEV criteria documents)
- Output characteristics

Ineligible

Bibliometric method

Informed peer-review method

Peer-review method

GEV final approval

Excellent

Good

Final

Fair

Result

Acceptable

Limited

Science
PHYSICS LETTERS A
The American Economic Review
nature
SOCIETY AND MENTAL HEALTH
INTELLECTUAL PATENTED PROPERTY
Sociological Theory

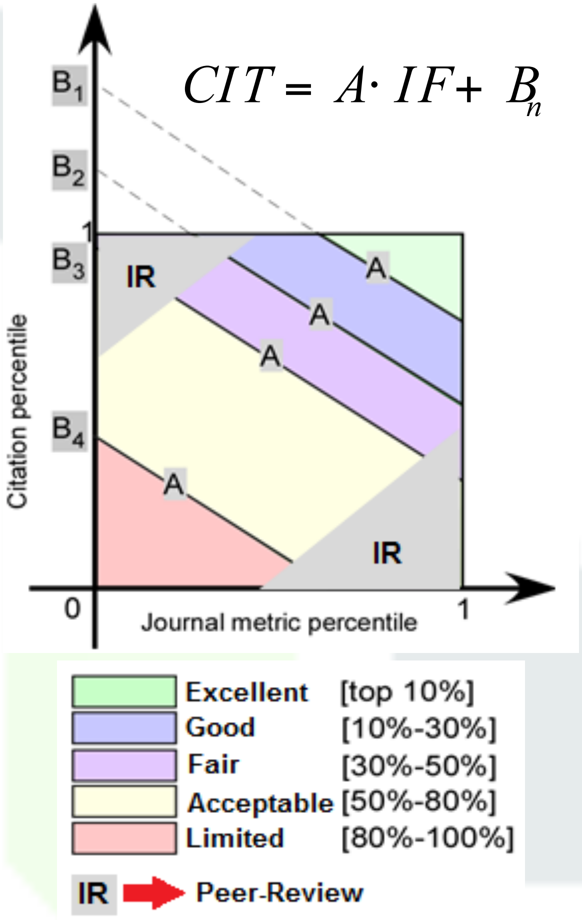
VQR 2011-2014: the bibliometric method

Bibliometric method

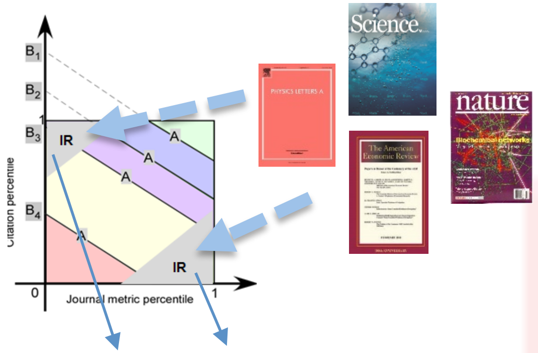
Based on a ***bibliometric algorithm*** which combines information from citations and Journal Impact Indicators

GEV 1*, 2, 3, 4, 5, 6, 7, 8b, 9, 11b, 13** evaluated articles published in journals indexed in WoS and Scopus databases.

GEV 1* adopted a slightly different bibliometric evaluation algorithm.
GEV 13** used a bibliometric algorithm significantly different from other bibliometric GEV, focusing on the publisher.



VQR 2011-2014: the peer-review method



Peer-review method

Peer-review carried out by (at least) two external experts (**reviewers**)

Independent experts chosen by **GEV members** in charge of the research output

GEV 8a, 10, 11a, 12, 14 assessed the totality of the research outputs using the peer-review method.

Evaluation was based on three multiple-choice questions, one for each of the following criteria: a) ***originality***, b) ***methodological rigor***, and c) ***attested or potential impact***. Each answer was assigned a score (1-10). The sum of the three scores was compared with four thresholds to generate a final classification into the same five classes as for the bibliometric algorithm.

VQR 2011-14: the peer review method

- In the following, we interpret a difference of more than two evaluation classes among the two reviewers as an indication of a high degree of uncertainty about the evaluation of a given output
- Hence, when comparing bibliometrics and peer review results, a discrepancy of two or three classes among the evaluation of the two reviewers is considered similar to the divergenge among the impact factor and the number of citations in the bibliometric analysis
- As a consequence, in order to compare the distribution of peer and bibliometric evaluations, an evaluation class denominated «Inconclusive peer» is created, similar to the class «Informed review» emerging from the bibliometric analysis

The sample of journal articles

- In order to evaluate how and to what extent bibliometric indicators are able to approximate the results of *peer-review*, we have randomly drawn a theoretical sample including 10% of all the products that underwent bibliometric evaluation.
- The **sample** has been stratified on the basis of GEV.
- Selected articles have been sent to *peer-review* using the same process that was used to select the experts in the VQR exercise.
- The number of articles effectively peer reviewed covers 9.3% of all the articles submitted to bibliometric evaluation (*next table*).
- No substantial selection biases due to the main characteristics of the papers (language, number of pages, bibliometric evaluation) has emerged from a post stratification analysis.

Articles evaluated with both peer-review and bibliometric methods

Scientific areas	Population	Empirical sample	% of the population
1- Mathematics and Computer Sciences	4.631	444	9,6
2 - Physics	10.182	1.008	9,9
3 - Chemistry	6.625	653	9,9
4 - Earth Sciences	3.953	388	9,8
5 - Biology	10.423	951	9,1
6 - Medicine	15.400	1.293	8,4
7 - Agricultural and veterinary sciences	6.354	630	9,9
8b - Civil Engineering	2.370	234	9,9
9 - Industrial and Information Engineering	9.930	890	9,0
11b – Psychology	1.801	175	9,7
13 - Economics and Statistics	5.490	498	9,1
Total	**77.159**	**7.164**	**9,3**

Bibliometric and peer evaluation

- The share of papers receiving the same evaluation according to both methods is approximately equal to 30%.
- Bibliometric evaluation is usually more generous than peer review, the latter concentrating most on central evaluation classes

Peer Bibliometric	Excellent	Good	Fair	Acceptable	Limited	Inconclusive Peer	Total
Excellent	10.5	20.1	4.5	0.6	0.0	5.4	41.1
Good	2.1	11.6	6.7	1.0	0.0	3.4	24.8
Fair	0.4	3.4	3.5	1.0	0.1	1.7	10.1
Acceptable	0.1	1.6	2.5	0.9	0.1	1.1	6.3
Limited	0.0	0.2	0.6	0.6	0.2	0.3	1.9
Informed review	0.8	6.1	4.8	1.1	0.1	2.7	15.7
Total	14.0	42.9	22.6	5.3	0.6	14.6	100.0

Rank association between bibliometric and peer evaluation

Rank correlation (Kendall’ Tau) among the two evaluation methods is equal to 0,39 on the whole sample, usually higher than the relationship among the two peer reviewers assigned to each paper

Scientific areas	Bibliometric and peer evaluation			First and second reviewer		
	Lower bound	tau-b	Upper bound	Lower bound	tau-b	Upper bound
Whole sample	0,374	0,393	0,412	0,273	0,293	0,313
1	0,231	0,317	0,403	0,172	0,257	0,342
2	0,335	0,383	0,431	0,186	0,242	0,298
3	0,239	0,309	0,379	0,119	0,190	0,261
4	0,255	0,338	0,421	0,155	0,243	0,332
5	0,320	0,374	0,427	0,189	0,247	0,305
6	0,361	0,406	0,450	0,215	0,261	0,307
7	0,281	0,355	0,428	0,199	0,267	0,334
8b	0,133	0,261	0,388	-0,067	0,059	0,185
9	0,252	0,313	0,374	0,182	0,182	0,245
11b	0,237	0,362	0,487	0,140	0,288	0,436
13	0,457	0,509	0,561	0,339	0,401	0,462

The use of bibliometric indicators in “predicting” peer- review results

- We assume that article quality may be well approximated by peer evaluation, and test whether bibliometric indicators may help in “predicting” peer review results.

- We estimate the following ordered probit model on the sample of articles that have been evaluated with both peer review and bibliometrics methods:

$$P(y_i = x) = F(b_1 cit_i + b_2 Impact_i + b_i z_i)$$

- ***F*** is the cumulative function of the Gaussian distribution
- $\boldsymbol{y_i}$ is the result of the peer evaluation for paper *i*.
- $\boldsymbol{P(y_i = x)}$ is the probability that evaluation is equal to *x* (*x* = 1; 0,7; 0,4; 0,1; 0).
- $\boldsymbol{cit_i}$ and $\boldsymbol{Impact_i}$ are the indicators for number of citations and journal impact for each paper.
- $\boldsymbol{z_i}$ is a set of characteristics that possibly influence evaluation results, including information related to the paper and the author: number of authors; national or international affiliation of the two reviewers; geographical location, academic role, age and gender of the author.

Estimation results

We restrict the analysis excluding papers presented in Area 13 (Economics and statistic), where a slightly different bibliometric algorithm was used
Overall, 6423 observations were included in the analysis.

	Coeff.	z	P>z
PAPER CHARACTERISTICS			
Number of authors	0.0004	7.91	0.00
Year of publication	-0.0072	-0.57	0.57
Mobility of the researcher (Hiring, promotion)	-0.0386	-1.00	0.32
First reviewer: Italian	0.0907	2.16	0.03
Second Reviewer: Italian	0.0792	1.88	0.06
Age	-0.0600	-3.55	0.00
Gender: Female	-0.0719	-2.36	0.02
INDIVIDUAL CHARACTERISTICS (control groups: GEV7, Veterinary sciences; PRO' researchers; North)			
GEV1: Mathematics	2.8806	24.71	0.00
GEV2: Physics	0.4305	6.81	0.00
GEV3: Chemistry	0.2592	4.14	0.00
GEV4: Earth Sciences	0.1182	1.54	0.12
GEV5: Biology	0.1436	2.43	0.02
GEV6: Medicine	-0.3179	-5.46	0.00
GEV8: Architecture	0.0386	0.45	0.66
GEV9: Engineering	0.0175	0.27	0.79
GEV11b: Psychology	0.2000	2.01	0.04
Full professor	-0.0234	-0.25	0.81
Associate professor	-0.1351	-1.50	0.13
Researcher (University)	-0.2835	-3.14	0.00
Center	-0.0443	-1.09	0.28
National	-0.2846	-3.18	0.00
South	-0.2436	-6.44	0.00
NA	0.1676	0.74	0.46
Journal Impact	0.0204	18.89	0.00
Citations	0.0123	14.20	0.00

Estimation results

- Better evaluations are obtained the higher the number of authors and the lower the age of the author
- No significant differences emerge for University' full and associate professors with respect to researchers in Public Research Organization (PRO), while University researchers obtain lower marks.
- Geographically, papers authored by researchers in the South get lower evaluations with respect to those coming from northern universities; also, researchers working in large PRO's with multiple locations in the Italian territory get lower marks with respect to those working in Universities and PRO's located in the north of the country.
- A slight negative gender effect for women also emerge from the analysis.
- On the other hand, no effects of being hired or promoted in the period considered or the year of publication of the paper is found in the analysis.

Conclusions and recommendations for future research

- The two bibliometric indicators used in the ANVUR algorithm, are both found to be strongly correlated with peer evaluation results, with the expected sign: once controlling for the main individual characteristics of the paper and its Authors, are both citations and journal impact found to play a significant role in ”predicting” article quality, as assessed by peer evaluation.

- When using peer review as the only method to evaluate scientific outcomes becomes unfeasible because of its high costs, we can hence conclude that the combined used of bibliometric indicators for citations and journal impact may provide a useful proxy to assess articles quality a good, albeit not unique and obviously amendable, proxy for peer review judgements. Results are robust to the exclusion of those variables from the analysis.

- Possible future research may imply testing for the correlation among peer review and bibliometric results also at the Institutional level (Traag and Waltman, 2017).

anvur

THANK YOU FOR YOUR ATTENTION!