UNIVERSITÀ DI SIENA 1240

USiena

Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise

This is the peer reviewed version of the following article:

Original:

Baccini, A., De Nicolao, G. (2016). Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. SCIENTOMETRICS, 108(3), 1651-1671 [10.1007/s11192-016-1929-y].

Availability:

This version is availablehttp://hdl.handle.net/11365/1005633 since 2018-09-20T16:19:21Z

Published:

DOI:10.1007/s11192-016-1929-y

Terms of use:

Open Access

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. Works made available under a Creative Commons license can be used according to the terms and conditions of said license.
For all terms of use and more information see the publisher's website.

(Article begins on next page)

24 February 2024

1. Introduction

context of the “Independent review of the role of metrics in research assessment and management”,

n article is published and so on. “This

–

”

. Italian VQR adopted a “dual system of evaluation”

in which both peer review and bibliometrics were considered. “In order to validate the use of the dual system”

“ experiment”

“In the complex … *a more than adequate concordance*

bibliometrics. This result fully justifies the choice made at VQR […] to use both

Final Report *Final Reports*

rest of the paper. Section 3 describes in some details ANVUR's experiment and re

2. The Italian research assessment exercise

Final Report *Area reports*

, the so called " ",

, called "sub GEV",

in a division between the so called "bibliometric areas", (Areas from 1 to 9 *hard*

the so called "no bibliometric areas" (Areas 10

(economics and statistics) was an exception: it was also classified as "non bibliometric" and the evaluation

Final Report *Area Reports*

stated that for each research field “a scale of values shared by the international scientific community” exists

- “scale of values
  international community”;
-
-
-

called “VQR distribution rule (20 50)”. The

Journal Citation Reports

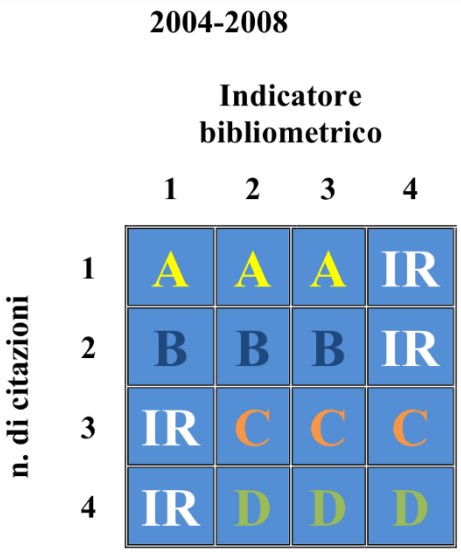

Figure 1.

viceversa

corresponding entry in the classification matrix was denoted by "IR" indicating that the article had to be submitted to a process called "informed peer review" (IR). For example, an article published in a journal with

ANVUR coined the expression “evaluative mix” to denote t

3. Comparing IR and bibliometrics

Final Report *Area reports*

Area 13 Report,

“ *”* .

Area 13 report

Final Report.

of the GEV, were asked to evaluate according “to their subjective

field”

were finally summarized in a final evaluation “based on algorithms specific for each Area”

Cohen’s kappa

Cohen’s kappa

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

p_o p_e

k $\kappa \leq 0$

The weights used in the calculation of Cohen’s kappa indicate the seriousness of the disagreement, by giving

test “is generally of little practical value, since a relatively low value of kappa can yield $k = 0.41$

areas corresponding to an administrative classification called “settori concorsuali”,

within a practical context”

Indeed the main problem is “how to maintain a consistent nomenclature when describing the relative strength of agreement associated with kappa statistics”

Table 1. Available guidelines for interpreting kappa values

Landis and Kock (1977)		**Altman (1991)**		**George and Mallery (2003)**	
K values	*Description*	*K values*	*Description*	*K values*	*Description*
Stemler and Tsai (2008)		**Fleiss (2003)**			
K values	*Description*	*K values*	*Description*		

“appropriate weights … associated with the qualitative evaluation”

lasting and consolidated stream of literature using Cohen’s kappa.

presents Cohen’s kappa

for Cohen’s kappa, by

Surprisingly enough, ANVUR's *Final Report* a as indicating "a *good* agreement for the whole sample and for each GEV" [italics added]; in the conclusion of the report, results are summarized by writing that there is "a more than adequate [in Italian "più che adeguata"] agreement luations done by adopting the peer review method and the bibliometric one" *verbatim* *Area Reports* Results of the experiment are also presented and commented as "giv evidence of a significant degree of concordance among peer review and bibliometric evaluations"

as "more than adequate"
"there is remarkable agreement between bibliometric and peer review evaluation"

that "informed peer review and bibliometric analysis produce similar evaluations" "fair to good agreement" in

The use made by ANVUR of expressions such as "good degree of agreement" and "more than adequate"

of agreement that can be described as "unacceptable", or alternatively as "poor" or "fair", for nearly all the

d inside Area 13, with an agreement that can be described as "acceptable" or alternatively as "fair to good" or "moderate".

---

even exchanged for agreement: "kappa is always statistically different from zero, showing that there is a fundamental agreement"

In the conclusion of the Area 9 report, that phrase is followed by the contradictory statement that: "The degree of ometric evaluations is moderate (in Italian: "moderato") in near all areas, while it results rather high (in Italian: "piuttosto elevato") for informatic engineering"

versions of the paper. They wrote "Since the most common scales to subjectively assess the value of kappa mention "adequate" and "fair to good", these are the terms we use in the paper." Really, the term "adequate" is not used in the

Table 2. Weighted kappas values for Areas and sub-areas.

Area 1 Mathematics and Informatics	**631**	**0,3176**	**0,3173**
Area 2 Physics	**1412**	**0,2302**	**0,2515**
Area 3 Chemistry	**927**	**0,2246**	**0,2296**
Area 4 Earth sciences	**458**	**0,2776**	**0,2985**
Area 5 Biology	**1310**	**0,3287**	**0,3453**
Area 6 Medicine	**1984**	**0,303**	**0,3351**
Area 7 Agricoltural and Veterinary sciences	**532**	**0,2776**	**0,3437**
Area 8 Civil engineering and Architecture	**225**	**0,1994**	**0,2261**

Area 9 Industrial and Information Engineering	**1130**	**0,1615**	**0,171**
Area 13 Economics and Statistics	**590**	**0,54**	**0,54**
All Areas	**9199**	**0,32**	**0,38**

Source. *Final Report* *Area Report*

4. A meta-analysis of the experiment

analyses were carried out, considering Cohen's

H0

from other areas. Recalling that Cohen's

m_j j *H0*

$k_j \sim N(\mu, \sigma^2/m_j)\ j = 1,2,\dots,n+1$

μ σ^2

$n = 9$ κ_{10} k_j $j = 1,2,\dots,9$

κ_{10}

m_{10}

κ_{10} *H0*

p

H0

H0

p

H0 1.2×10^{-4}

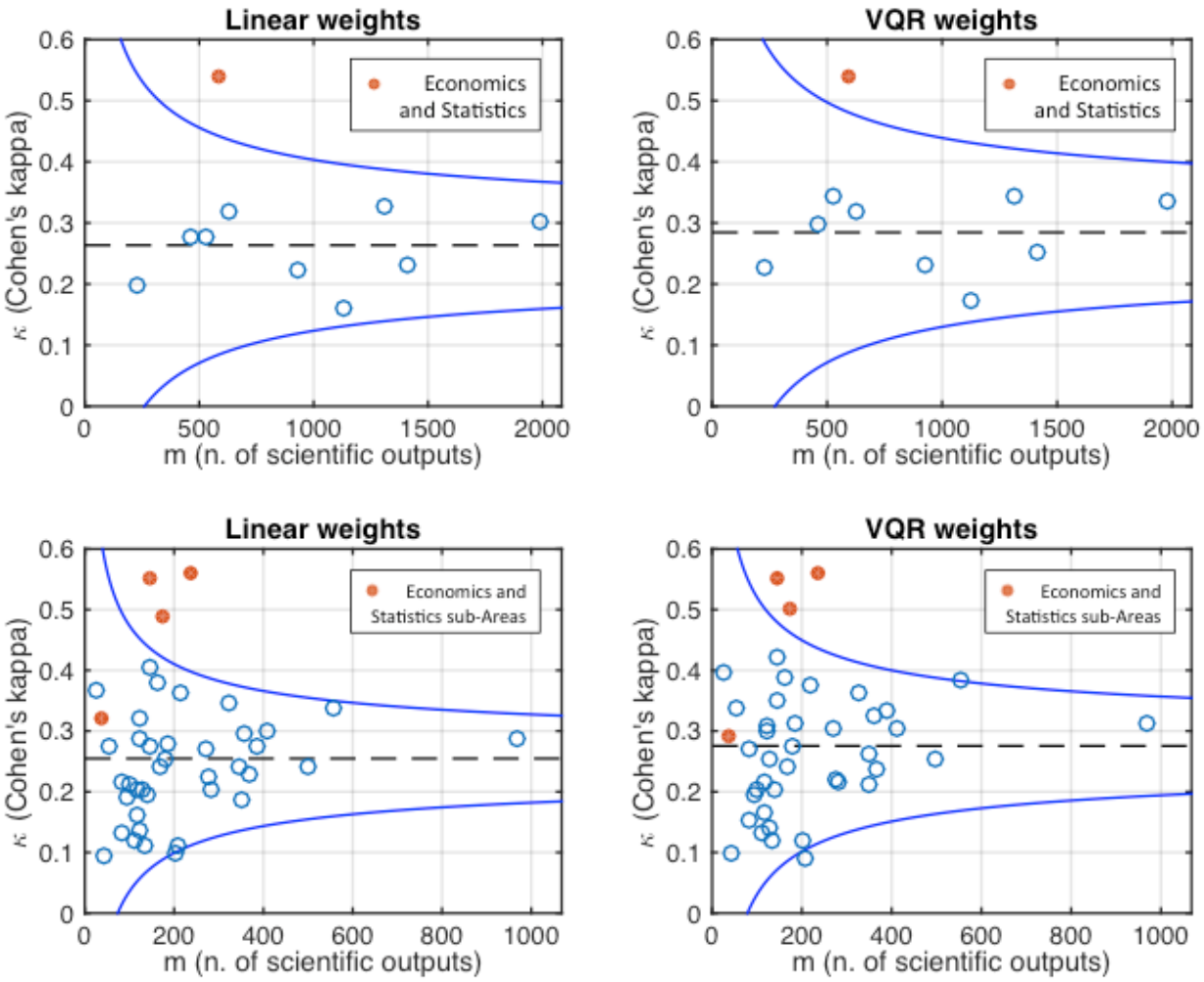

Figure 2. *m* κ *m*

ose Cohen's κ. Cohen's

$(\kappa = 0.09)$

Table 3. *p*-values for Area 13 and its sub-areas.

Sub-areas	
$p < 0.05$;	$p < 0.01$ $p < 0.001$

5. A self-fulfilling experiment

ANVUR's *Final Report*

authors'

B of the Area Report: "The sample selection shall take accoun

" (p.64).

"The

[five years impact factor] and AIS [Article influence score]"

article's bibliometric evaluation: he just had to check the journal ranking. So in Area 13, not only refere

scores. In Area 13 the protocol was completely different. When the two referee's reports were comunicated to the two GEV members in charge of the considered article, they formed a "consensus group" which di decided the final evaluation of the article, by considering the referee's reports as simple information for their

ANVUR's *Final Report*

ed in the appendix A of the Area 13 Report: "The

between [the two referee's reports], the [final score] is not simply the average of [the referee'

the peer review process)" The work of the consensus groups is described as follows: "The Consensus

and the Consensus Group competences."
competences of the two referees, and gave "more importance to the most expert referee in the research field". (Area

assessment exercise the notion of "informed peer review" individuates at least two very different pro

13 can be interpreted as indicating a "fair to good agreement" between the evaluation based on the journal ranking

subarea of “ history”

area “

managed to give importance, in different measures, to external referees’ opinions and, consequently most expert referee’s point of view”

---

as pertaining to “economic history” were evaluated by the panelist of the sub area “economics”. I

6. Concluding remarks

statistical technique (the weighted Cohen's kappa). ANVUR official reports interpreted these results indicating an overall "good" or "more than adequate" agreement between IR and bibliometrics. This result

has to be interpreted as "unacceptable", "poor" or in a few cases as, at most, "fair". The only notable

authors'

two referees'

Ceteris paribus

mix of instruments used, but also by the intrinsic "lower quality" of research outputs which were evaluated

: “results of the analysis relative to the degree of concordance …
may be considered to validate the general approach of combining peer review and bibliometric methods”
The conclusion reached by ANVUR, according to which the “more than adequate”
agreement between bibliometric and informed peer review “fully justifies” the use of both techniques of

A fortiori

the two techniques cannot be considered as “substitute”, as sustained by

appears to be well founded: “the agencies that run these evaluations could feel confident about using

performed informed peer review”

lysis performed in the context of the “Independent
review of the role of metrics in research assessment and management”

Appendix

–

the 590 articles products can be treated as they were evaluated by two referees. Let's s

Source:

I voti all'università. La Valutazione della qualità della ricerca in Italia

Journal of the Association for Information Science and Technology

Research Evaluation, 13

PLoS ONE, 4

Practical statistics for medical research

– Research Evaluation, 24

Statistica & Società, 3

http://www.roars.it/online/lo-strano-caso-delle-concordanze-della-vqr/

Baccini, A. (2016). Napoléon et l'évaluation bibliométrique de la recherche. Considérations sur la réforme de l'université et sur l'action de l'agence national d'évaluation en Italie. *Canadian Journal of Information and Library Science-Revue Canadienne des Sciences de l'Information et de Bibliotheconomie*

Annals of Oncology, 14

Department of Economics DEMB

ReCent WP

IZA Discussion paper

CEPR Discussion papers

CSEF working papers

www.voxeu.org

Research Policy, 44

MPRA (Munich Personal REPEc Archive)

Educational and Psychological Measurement, 20

Psychological Bulletin, 70

De Nicolao, G. (2014). VQR da buttare? Persino ANVUR cestina i voti usati per l'assegnazione FFO 2013. *http://www.roars.it/online/vqr-da-buttare-persino-anvur-cestina-i-voti-usati-per-lassegnazione-ffo-2013/*

Statistical Methods for Rates and Proportions

SPSS for Windows Step By Step: A simple Guide and Reference

Journal of the American Society for Information Science, 34

Biometrics, 33

The Research Assessment Exercise, the state and the dominance of mainstream economics in British universities

BioScience, 58

McNay, I. (2011). Research assessment: work in progress, or 'la lutta continua' In M. Saunders, P. Trowler, & *Reconceptualising Evaluation in Higher Education The Practice Turn*

Scientometrics, 102

Reseach Policy, 27

Handbook of Parametric and Nonparametric Statistical Procedures

Stat Med, 24

Best practices in quantitative methods

analysis of Cohen's kappa. *Health Services and Outcomes Research Methodology, 11*

Scientometrics, 67