AISA
Associazione italiana per la promozione della scienza aperta

Terzo convegno annuale

Scienza aperta e integrità della ricerca

ANVUR: i dati chiusi della bibliometria di stato

Alberto Baccini
Università di Siena
Giuseppe De Nicolao
Università di Pavia

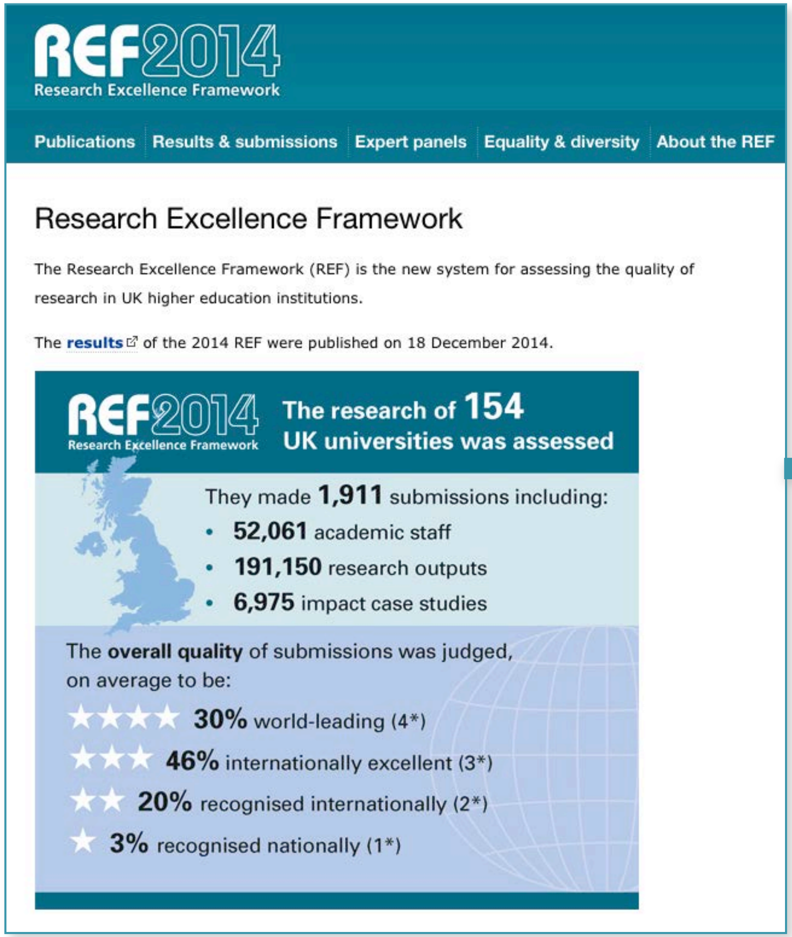
ROARS
Return On Academic ReSearch

UNIVERSITÀ DI PAVIA

Sommario

1. Valutazione della ricerca: lo stato dell’arte nel 2011
2. VQR, la via italiana alla valutazione della ricerca
3. Cronaca di un esperimento annunciato
4. Bibliometrics vs peer review: do they agree?
5. Concordanza o fallacia statistica?
6. Dati chiusi, concordanza non replicabile
7. Conclusioni

1. Valutazione della ricerca: lo stato dell’arte nel 2011



Research Excellence Framework

The Research Excellence Framework (REF) is the new system for assessing the quality of research in UK higher education institutions.

The **results** of the 2014 REF were published on 18 December 2014.

REF2014 Research Excellence Framework

The research of 154 UK universities was assessed

They made **1,911** submissions including:

- **52,061** academic staff
- **191,150** research outputs
- **6,975** impact case studies

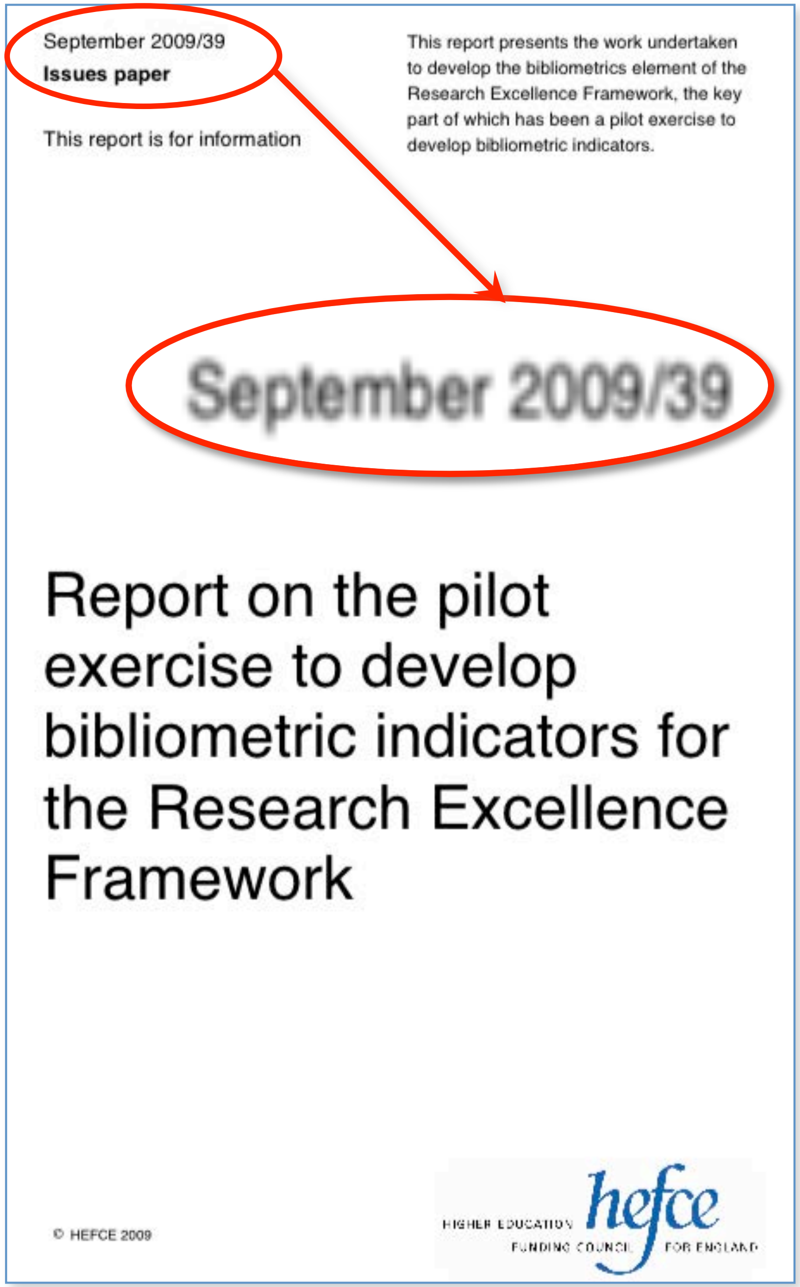
The **overall quality** of submissions was judged, on average to be:

★★★★ **30%** world-leading (4*)

★★★ **46%** internationally excellent (3*)

★★ **20%** recognised internationally (2*)

★ **3%** recognised nationally (1*)



September 2009/39

Issues paper

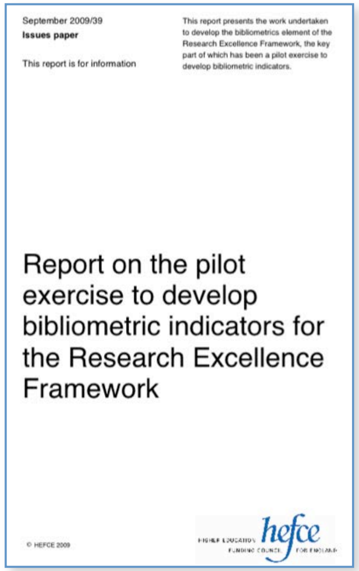
This report is for information

This report presents the work undertaken to develop the bibliometrics element of the Research Excellence Framework, the key part of which has been a pilot exercise to develop bibliometric indicators.

September 2009/39

Report on the pilot exercise to develop bibliometric indicators for the Research Excellence Framework

© HEFCE 2009



September 2009/39

Issues paper

This report is for information

This report presents the work undertaken to develop the bibliometrics element of the Research Excellence Framework, the key part of which has been a pilot exercise to develop bibliometric indicators.

Report on the pilot exercise to develop bibliometric indicators for the Research Excellence Framework

© HEFCE 2009

hefce HIGHER EDUCATION FUNDING COUNCIL FOR ENGLAND

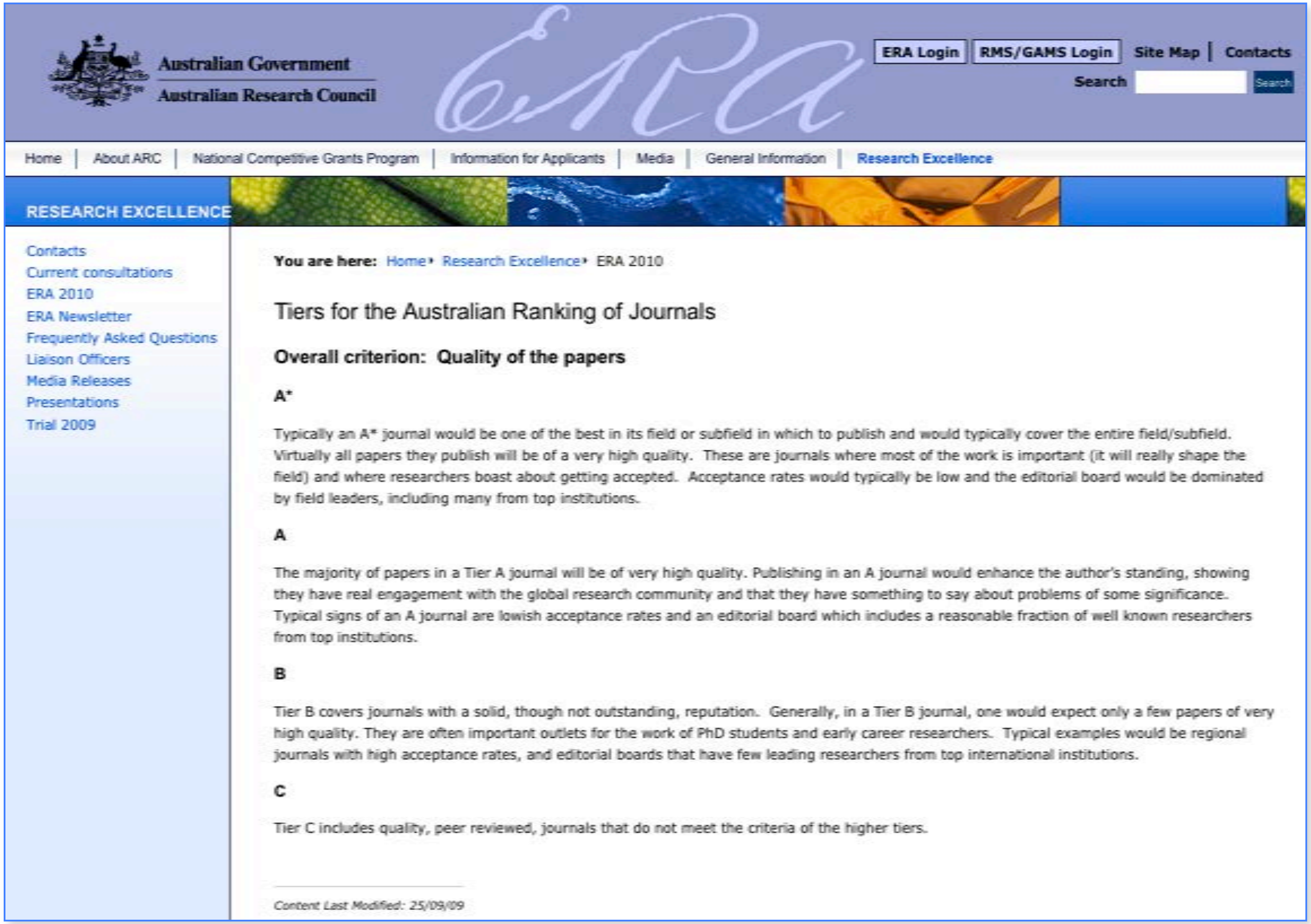
Key points

8. Bibliometrics are not sufficiently robust at this stage to be used formulaically or to replace expert review in the REF. However there is considerable scope for citation information to be used to inform expert review.

The Australian Way

the-australian-way.de

ERA 2010: Ranking of Journals



You are here: Home › Research Excellence › ERA 2010

Tiers for the Australian Ranking of Journals

Overall criterion: Quality of the papers

A*

Typically an A* journal would be one of the best in its field or subfield in which to publish and would typically cover the entire field/subfield. Virtually all papers they publish will be of a very high quality. These are journals where most of the work is important (it will really shape the field) and where researchers boast about getting accepted. Acceptance rates would typically be low and the editorial board would be dominated by field leaders, including many from top institutions.

A

The majority of papers in a Tier A journal will be of very high quality. Publishing in an A journal would enhance the author's standing, showing they have real engagement with the global research community and that they have something to say about problems of some significance. Typical signs of an A journal are lowish acceptance rates and an editorial board which includes a reasonable fraction of well known researchers from top institutions.

B

Tier B covers journals with a solid, though not outstanding, reputation. Generally, in a Tier B journal, one would expect only a few papers of very high quality. They are often important outlets for the work of PhD students and early career researchers. Typical examples would be regional journals with high acceptance rates, and editorial boards that have few leading researchers from top international institutions.

C

Tier C includes quality, peer reviewed, journals that do not meet the criteria of the higher tiers.

Content Last Modified: 25/09/09

30 maggio 2011

Kim Carr: «*There is clear and consistent evidence that the rankings were being deployed inappropriately within some quarters of the sector, in ways that could produce harmful outcomes [...]. [...] **the removal of the ranks** and the provision of the publication profile will ensure they will be used descriptively rather than prescriptively.*»

Kim Carr, the Australian Minister for Innovation, Industry, Science and Research

House of Commons

Science and Technology Committee

Peer review in scientific publications

Eighth Report of Session 2010–12

Volume I: Report, together with formal minutes, oral and written evidence

Additional written evidence is contained in Volume II, available on the Committee website at www.parliament.uk/science

Ordered by the House of Commons to be printed 18 July 2011

David Sweeney [Director HEFCE]: «*it is an underpinning element in the exercise that **journal impact factors will not be used**. I think we were very interested to see that in Australia, where they conceived an exercise that was heavily dependent on journal rankings, after carrying out the first exercise, they decided that alternative ways of assessing quality*»

ICIAM

International Mathematical Union (IMU)

ims

Joint Committee on Quantitative Assessment of Research

Citation Statistics

A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS)

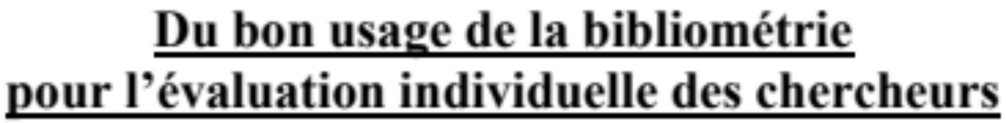
Corrected version, 6/12/08

"The idea that research assessment must be done using "simple and objective" methods is increasingly prevalent today. The "simple and objective" methods are broadly interpreted as bibliometrics, that is, citation data and the statistics derived from them. There is a belief that citation statistics are inherently more accurate because they substitute simple numbers for complex judgments, and hence overcome the possible subjectivity of peer review. But ***this belief is unfounded****."*

INSTITUT DE FRANCE

Académie des sciences

17 gennaio 2011

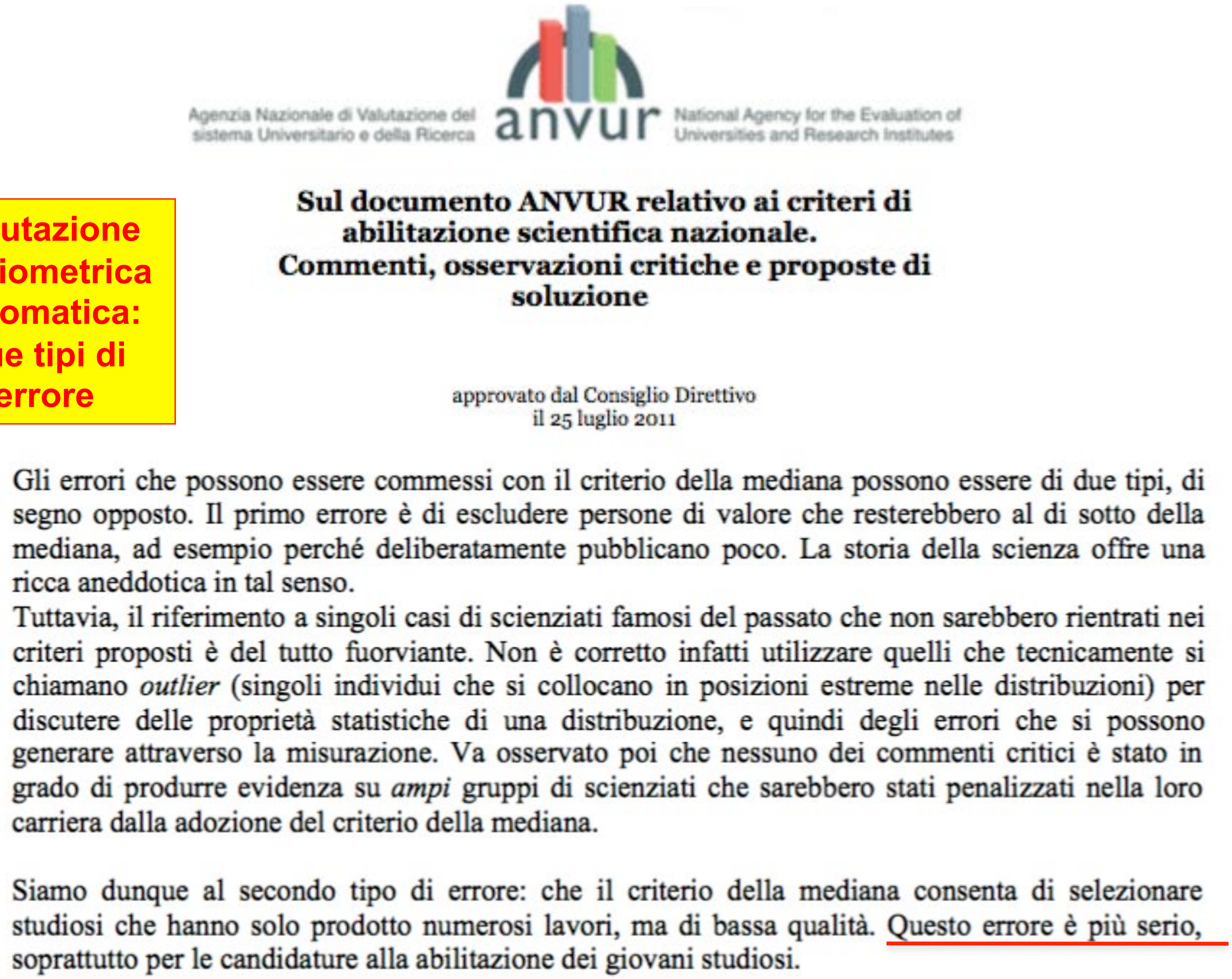


Du bon usage de la bibliométrie
pour l'évaluation individuelle des chercheurs

"Any bibliometric evaluation should be tightly associated to a close examination of a researcher's work, in particular to evaluate its originality, an element that cannot be assessed through a bibliometric study."

2. VQR, la via italiana alla valutazione della ricerca

Valutazione bibliometrica automatica: due tipi di errore



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

Sul documento ANVUR relativo ai criteri di abilitazione scientifica nazionale. Commenti, osservazioni critiche e proposte di soluzione

approvato dal Consiglio Direttivo
il 25 luglio 2011

1 Gli errori che possono essere commessi con il criterio della mediana possono essere di due tipi, di segno opposto. Il primo errore è di escludere persone di valore che resterebbero al di sotto della mediana, ad esempio perché deliberatamente pubblicano poco. La storia della scienza offre una ricca aneddotica in tal senso.
Tuttavia, il riferimento a singoli casi di scienziati famosi del passato che non sarebbero rientrati nei criteri proposti è del tutto fuorviante. Non è corretto infatti utilizzare quelli che tecnicamente si chiamano *outlier* (singoli individui che si collocano in posizioni estreme nelle distribuzioni) per discutere delle proprietà statistiche di una distribuzione, e quindi degli errori che si possono generare attraverso la misurazione. Va osservato poi che nessuno dei commenti critici è stato in grado di produrre evidenza su *ampi* gruppi di scienziati che sarebbero stati penalizzati nella loro carriera dalla adozione del criterio della mediana.

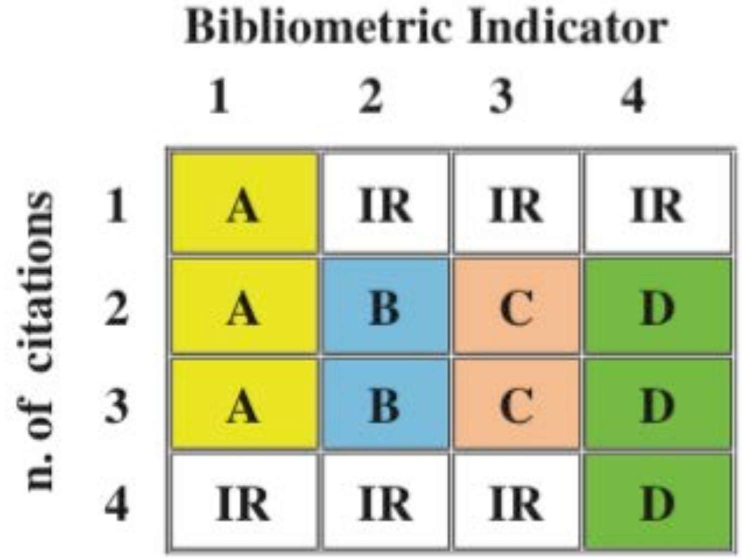
2 Siamo dunque al secondo tipo di errore: che il criterio della mediana consenta di selezionare studiosi che hanno solo prodotto numerosi lavori, ma di bassa qualità. Questo errore è più serio, soprattutto per le candidature alla abilitazione dei giovani studiosi.

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

“*gli errori che possono essere commessi con il criterio della mediana possono essere di due tipi, di segno opposto****. Il primo errore è di escludere persone di valore*** [...] *Siamo dunque al secondo tipo di errore: che il criterio della mediana consenta di selezionare studiosi che hanno solo prodotto numerosi lavori, ma di bassa qualità.* ***Questo errore è più serio***”

Il “mix valutativo” della VQR 2004-2010

- Inedito metodo bibliometrico:



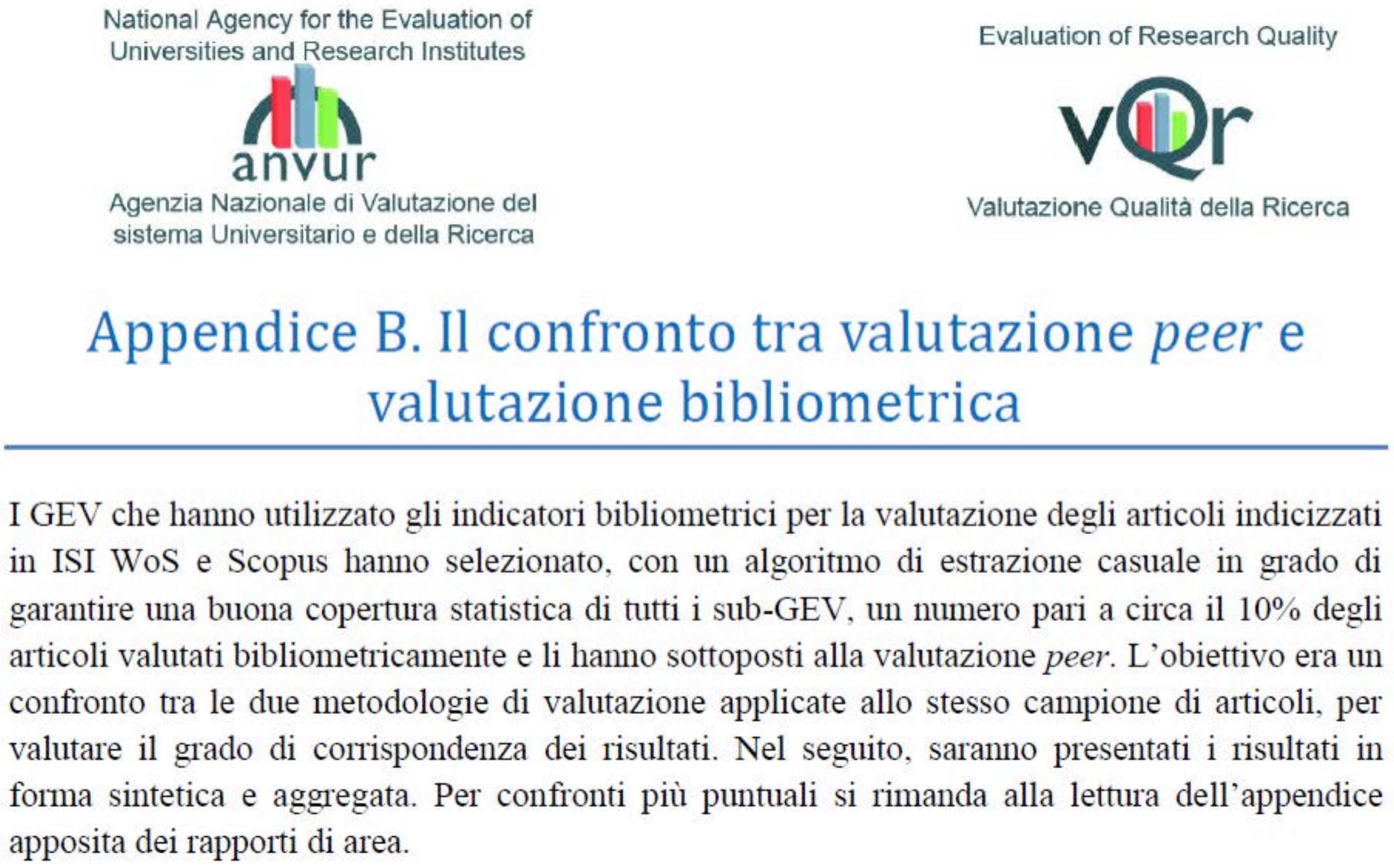
Bibliometric Indicator

n. of citations	1	2	3	4
1	A	IR	IR	IR
2	A	B	C	D
3	A	B	C	D
4	IR	IR	IR	D

Figure 2. The Bibliometric matrix.
Source: ANVUR.

- Si usano insieme peer review e bibliometria

Ma è lecito mescolare peer review e bibliometria?



National Agency for the Evaluation of Universities and Research Institutes

anvur

Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca

Evaluation of Research Quality

vQr

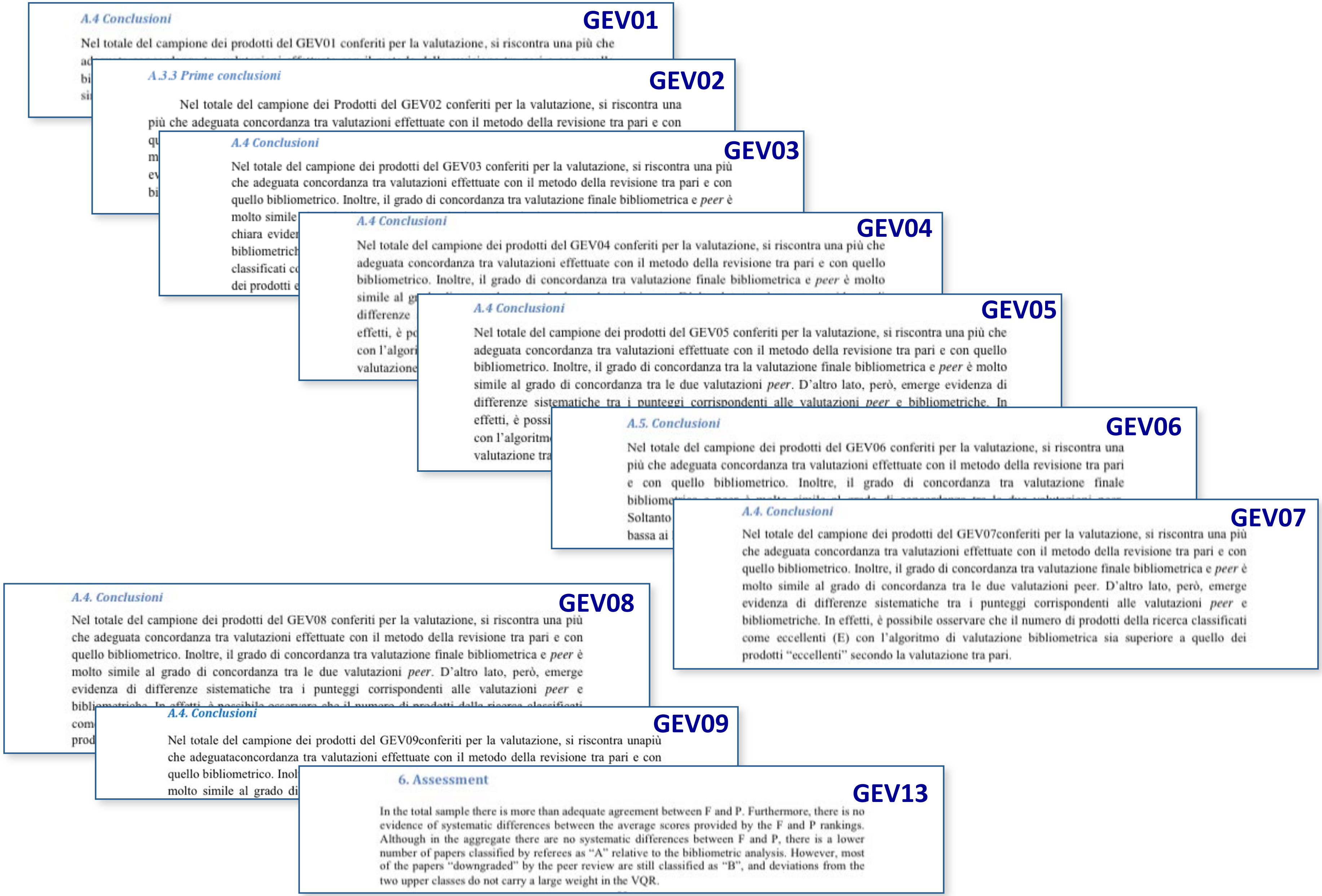
Valutazione Qualità della Ricerca

Appendice B. Il confronto tra valutazione *peer* e valutazione bibliometrica

I GEV che hanno utilizzato gli indicatori bibliometrici per la valutazione degli articoli indicizzati in ISI WoS e Scopus hanno selezionato, con un algoritmo di estrazione casuale in grado di garantire una buona copertura statistica di tutti i sub-GEV, un numero pari a circa il 10% degli articoli valutati bibliometricamente e li hanno sottoposti alla valutazione *peer*. L'obiettivo era un confronto tra le due metodologie di valutazione applicate allo stesso campione di articoli, per valutare il grado di corrispondenza dei risultati. Nel seguito, saranno presentati i risultati in forma sintetica e aggregata. Per confronti più puntuali si rimanda alla lettura dell'appendice apposita dei rapporti di area.

3. Cronaca di un esperimento annunciato

Conclusioni tutte uguali



GEV01

A.4 Conclusioni

Nel totale del campione dei prodotti del GEV01 conferiti per la valutazione, si riscontra una più che
ad
bi
si

GEV02

A.3.3 Prime conclusioni

Nel totale del campione dei Prodotti del GEV02 conferiti per la valutazione, si riscontra una più che adeguata concordanza tra valutazioni effettuate con il metodo della revisione tra pari e con
qu
m
ev
bi

GEV03

A.4 Conclusioni

Nel totale del campione dei prodotti del GEV03 conferiti per la valutazione, si riscontra una più che adeguata concordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inoltre, il grado di concordanza tra valutazione finale bibliometrica e *peer* è
molto simile
chiara evider
bibliometrich
classificati co
dei prodotti e

GEV04

A.4 Conclusioni

Nel totale del campione dei prodotti del GEV04 conferiti per la valutazione, si riscontra una più che adeguata concordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inoltre, il grado di concordanza tra valutazione finale bibliometrica e *peer* è molto
simile al gr
differenze
effetti, è po
con l'algori
valutazione

GEV05

A.4 Conclusioni

Nel totale del campione dei prodotti del GEV05 conferiti per la valutazione, si riscontra una più che adeguata concordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inoltre, il grado di concordanza tra la valutazione finale bibliometrica e *peer* è molto simile al grado di concordanza tra le due valutazioni *peer*. D'altro lato, però, emerge evidenza di differenze sistematiche tra i punteggi corrispondenti alle valutazioni *peer* e bibliometriche. In
effetti, è possi
con l'algoritm
valutazione tra

GEV06

A.5. Conclusioni

Nel totale del campione dei prodotti del GEV06 conferiti per la valutazione, si riscontra una più che adeguata concordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inoltre, il grado di concordanza tra valutazione finale
bibliome
Soltanto
bassa ai

GEV07

A.4. Conclusioni

Nel totale del campione dei prodotti del GEV07conferiti per la valutazione, si riscontra una più che adeguata concordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inoltre, il grado di concordanza tra valutazione finale bibliometrica e *peer* è molto simile al grado di concordanza tra le due valutazioni peer. D'altro lato, però, emerge evidenza di differenze sistematiche tra i punteggi corrispondenti alle valutazioni *peer* e bibliometriche. In effetti, è possibile osservare che il numero di prodotti della ricerca classificati come eccellenti (E) con l'algoritmo di valutazione bibliometrica sia superiore a quello dei prodotti "eccellenti" secondo la valutazione tra pari.

GEV08

A.4. Conclusioni

Nel totale del campione dei prodotti del GEV08 conferiti per la valutazione, si riscontra una più che adeguata concordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inoltre, il grado di concordanza tra valutazione finale bibliometrica e *peer* è molto simile al grado di concordanza tra le due valutazioni *peer*. D'altro lato, però, emerge evidenza di differenze sistematiche tra i punteggi corrispondenti alle valutazioni *peer* e
bibli
com
prod

GEV09

A.4. Conclusioni

Nel totale del campione dei prodotti del GEV09conferiti per la valutazione, si riscontra unapiù che adeguataconcordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inol
molto simile al grado di

GEV13

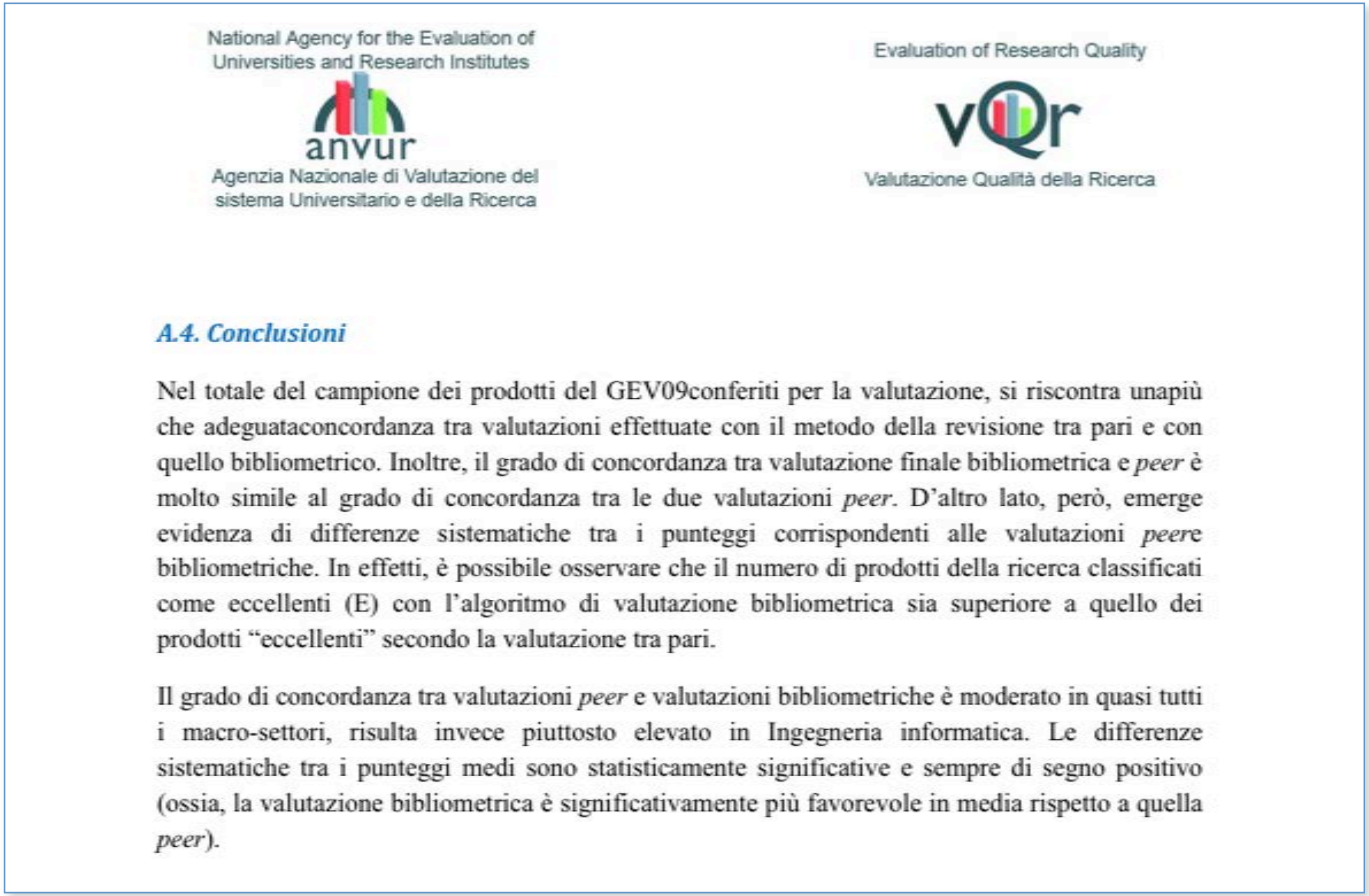
6. Assessment

In the total sample there is more than adequate agreement between F and P. Furthermore, there is no evidence of systematic differences between the average scores provided by the F and P rankings. Although in the aggregate there are no systematic differences between F and P, there is a lower number of papers classified by referees as "A" relative to the bibliometric analysis. However, most of the papers "downgraded" by the peer review are still classified as "B", and deviations from the two upper classes do not carry a large weight in the VQR.

Conclusioni tutte uguali

“Nel totale del campione dei prodotti del GEV_X conferiti per la valutazione, si riscontra una più che adeguata concordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico.”

Conclusioni tutte uguali … o quasi



National Agency for the Evaluation of
Universities and Research Institutes

anvur

Agenzia Nazionale di Valutazione del
sistema Universitario e della Ricerca

Evaluation of Research Quality

vQr

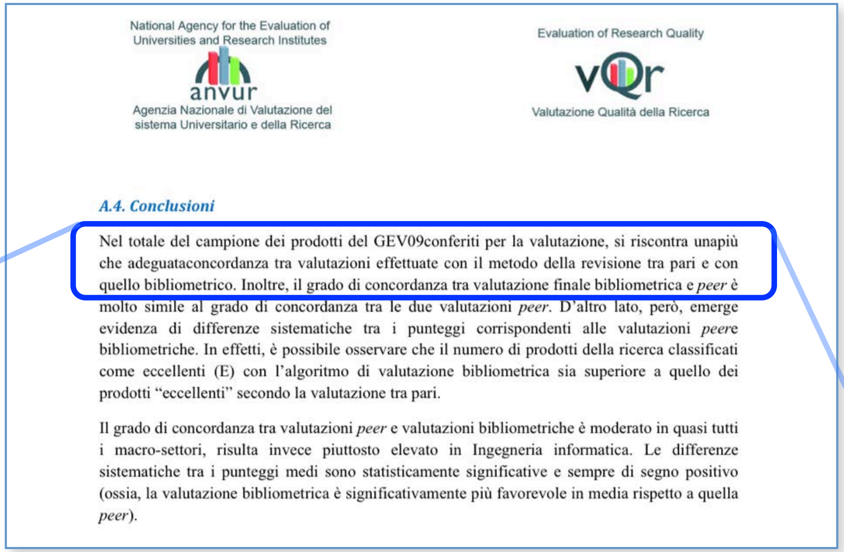
Valutazione Qualità della Ricerca

A.4. Conclusioni

Nel totale del campione dei prodotti del GEV09conferiti per la valutazione, si riscontra unapiù che adeguataconcordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inoltre, il grado di concordanza tra valutazione finale bibliometrica e *peer* è molto simile al grado di concordanza tra le due valutazioni *peer*. D'altro lato, però, emerge evidenza di differenze sistematiche tra i punteggi corrispondenti alle valutazioni *peer*e bibliometriche. In effetti, è possibile osservare che il numero di prodotti della ricerca classificati come eccellenti (E) con l'algoritmo di valutazione bibliometrica sia superiore a quello dei prodotti "eccellenti" secondo la valutazione tra pari.

Il grado di concordanza tra valutazioni *peer* e valutazioni bibliometriche è moderato in quasi tutti i macro-settori, risulta invece piuttosto elevato in Ingegneria informatica. Le differenze sistematiche tra i punteggi medi sono statisticamente significative e sempre di segno positivo (ossia, la valutazione bibliometrica è significativamente più favorevole in media rispetto a quella *peer*).

Facciamo uno zoom sul Rapporto di Area 09



National Agency for the Evaluation of Universities and Research Institutes

anvur

Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca

Evaluation of Research Quality

vQr

Valutazione Qualità della Ricerca

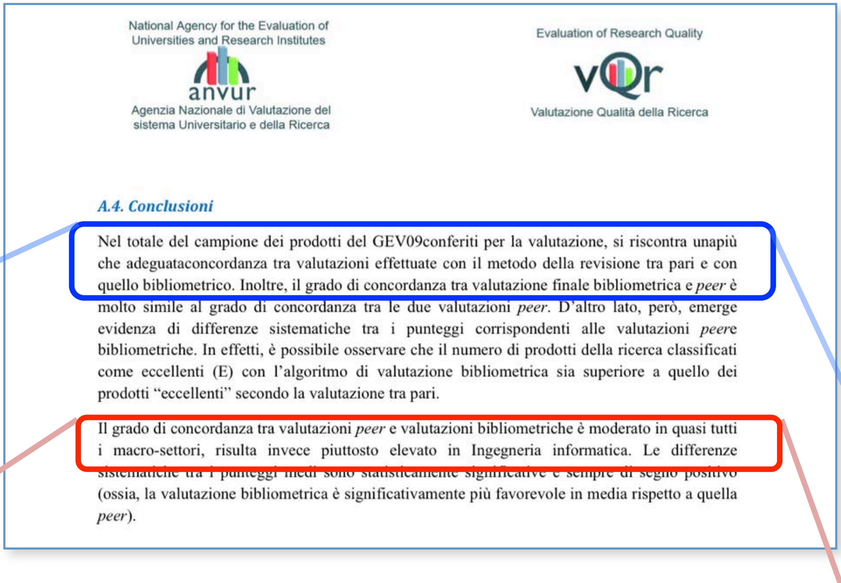
A.4. Conclusioni

Nel totale del campione dei prodotti del GEV09conferiti per la valutazione, si riscontra unapiù che adeguataconcordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inoltre, il grado di concordanza tra valutazione finale bibliometrica e *peer* è molto simile al grado di concordanza tra le due valutazioni *peer*. D'altro lato, però, emerge evidenza di differenze sistematiche tra i punteggi corrispondenti alle valutazioni *peer*e bibliometriche. In effetti, è possibile osservare che il numero di prodotti della ricerca classificati come eccellenti (E) con l'algoritmo di valutazione bibliometrica sia superiore a quello dei prodotti "eccellenti" secondo la valutazione tra pari.

Il grado di concordanza tra valutazioni *peer* e valutazioni bibliometriche è moderato in quasi tutti i macro-settori, risulta invece piuttosto elevato in Ingegneria informatica. Le differenze sistematiche tra i punteggi medi sono statisticamente significative e sempre di segno positivo (ossia, la valutazione bibliometrica è significativamente più favorevole in media rispetto a quella *peer*).

Nel totale del campione dei prodotti del GEV09conferiti per la valutazione, si riscontra unapiù che adeguataconcordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inoltre, il grado di concordanza tra valutazione finale bibliometrica e *peer* è

Rapporto di Area 09



National Agency for the Evaluation of Universities and Research Institutes

anvur

Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca

Evaluation of Research Quality

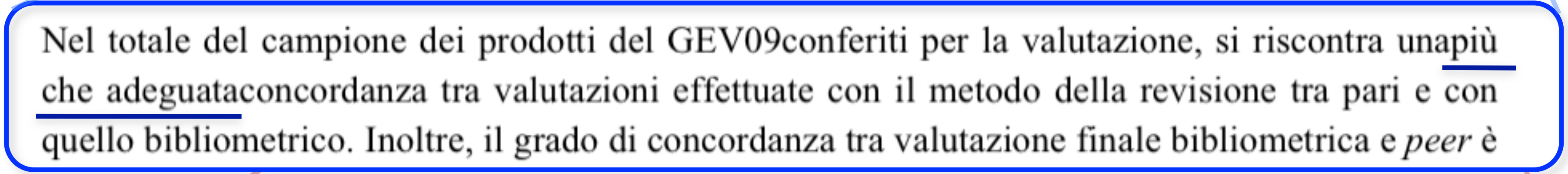
vQr

Valutazione Qualità della Ricerca

A.4. Conclusioni

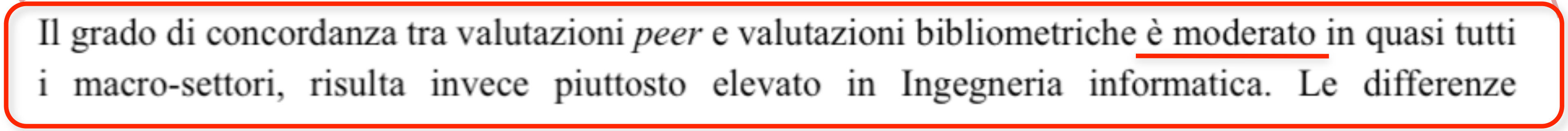
Nel totale del campione dei prodotti del GEV09conferiti per la valutazione, si riscontra unapiù che adeguataconcordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inoltre, il grado di concordanza tra valutazione finale bibliometrica e *peer* è molto simile al grado di concordanza tra le due valutazioni *peer*. D'altro lato, però, emerge evidenza di differenze sistematiche tra i punteggi corrispondenti alle valutazioni *peer*e bibliometriche. In effetti, è possibile osservare che il numero di prodotti della ricerca classificati come eccellenti (E) con l'algoritmo di valutazione bibliometrica sia superiore a quello dei prodotti "eccellenti" secondo la valutazione tra pari.

Il grado di concordanza tra valutazioni *peer* e valutazioni bibliometriche è moderato in quasi tutti i macro-settori, risulta invece piuttosto elevato in Ingegneria informatica. Le differenze sistematiche tra i punteggi medi sono statisticamente significative e sempre di segno positivo (ossia, la valutazione bibliometrica è significativamente più favorevole in media rispetto a quella *peer*).



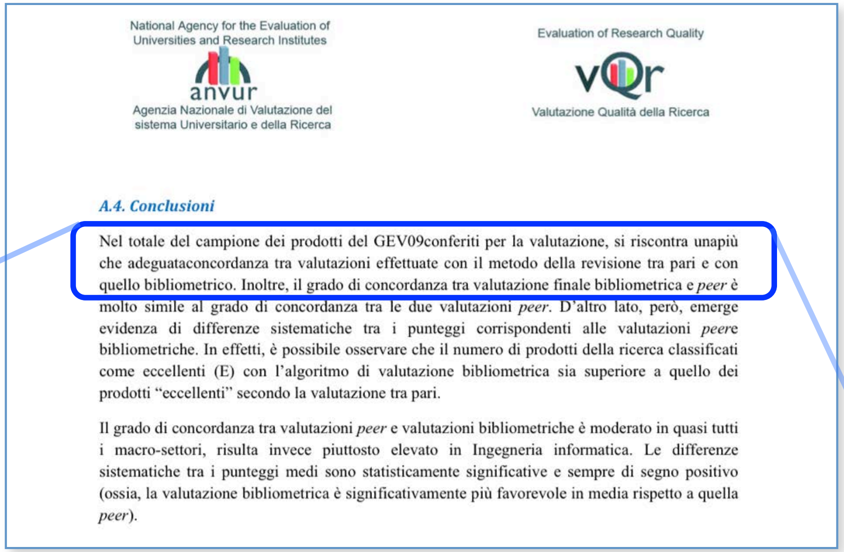
Nel totale del campione dei prodotti del GEV09conferiti per la valutazione, si riscontra unapiù che adeguataconcordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inoltre, il grado di concordanza tra valutazione finale bibliometrica e *peer* è

ma la concordanza è **più che adeguata** o **moderata**?



Il grado di concordanza tra valutazioni *peer* e valutazioni bibliometriche è moderato in quasi tutti i macro-settori, risulta invece piuttosto elevato in Ingegneria informatica. Le differenze

Facciamo uno zoom sul Rapporto di Area 09



National Agency for the Evaluation of Universities and Research Institutes

anvur

Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca

Evaluation of Research Quality

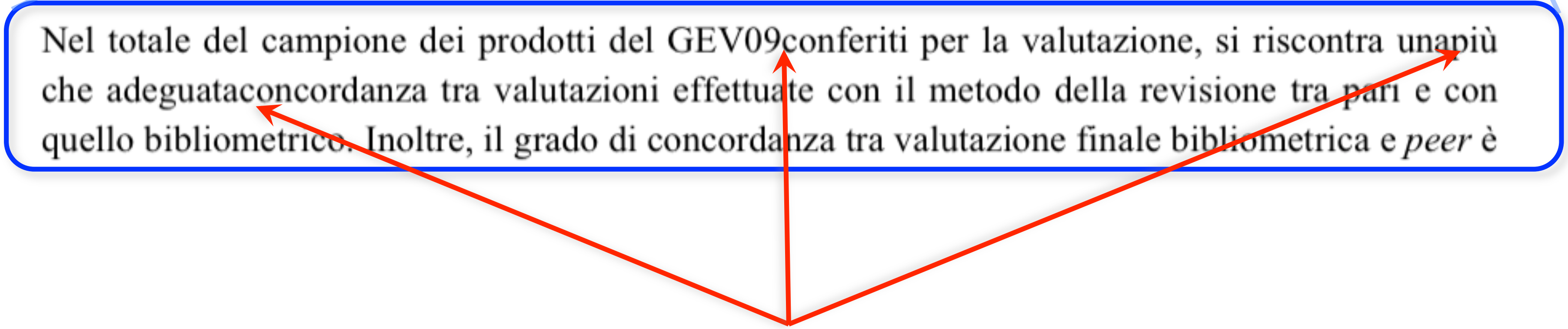
vqr

Valutazione Qualità della Ricerca

A.4. Conclusioni

Nel totale del campione dei prodotti del GEV09conferiti per la valutazione, si riscontra unapiù che adeguataconcordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inoltre, il grado di concordanza tra valutazione finale bibliometrica e *peer* è molto simile al grado di concordanza tra le due valutazioni *peer*. D'altro lato, però, emerge evidenza di differenze sistematiche tra i punteggi corrispondenti alle valutazioni *peere* bibliometriche. In effetti, è possibile osservare che il numero di prodotti della ricerca classificati come eccellenti (E) con l'algoritmo di valutazione bibliometrica sia superiore a quello dei prodotti "eccellenti" secondo la valutazione tra pari.

Il grado di concordanza tra valutazioni *peer* e valutazioni bibliometriche è moderato in quasi tutti i macro-settori, risulta invece piuttosto elevato in Ingegneria informatica. Le differenze sistematiche tra i punteggi medi sono statisticamente significative e sempre di segno positivo (ossia, la valutazione bibliometrica è significativamente più favorevole in media rispetto a quella *peer*).

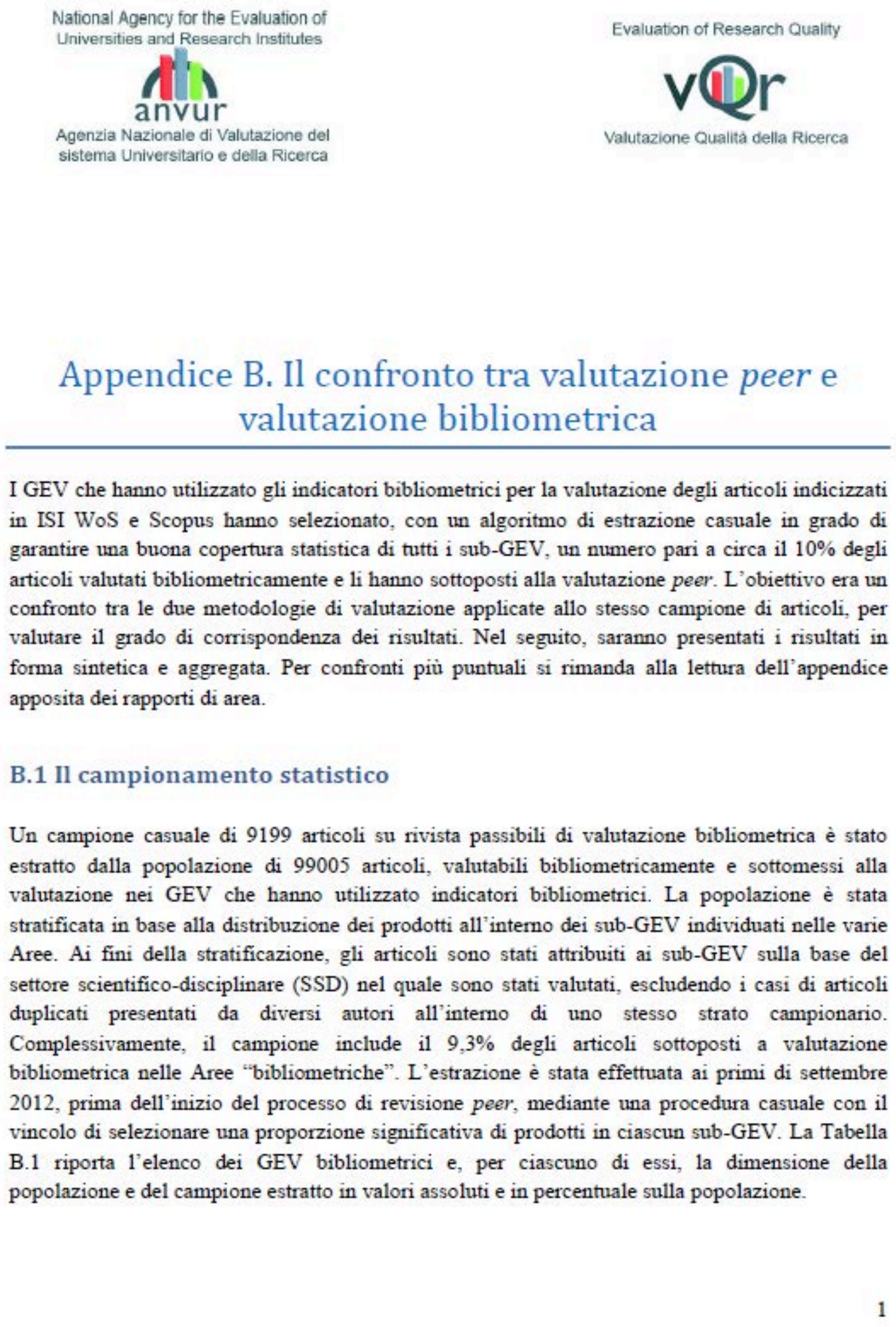


Nel totale del campione dei prodotti del GEV09conferiti per la valutazione, si riscontra unapiù che adeguataconcordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inoltre, il grado di concordanza tra valutazione finale bibliometrica e *peer* è

Mancano degli spazi.

Non è che il rapporto dell'area 09 (**quella con la concordanza peggiore**), ha subito una correzione "last minute" per uniformarlo agli altri rapporti, con una sostituzione che richiedeva più caratteri?

Un rapporto, molti working papers e anche un articolo scientifico



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

Evaluation of Research Quality
vQr
Valutazione Qualità della Ricerca

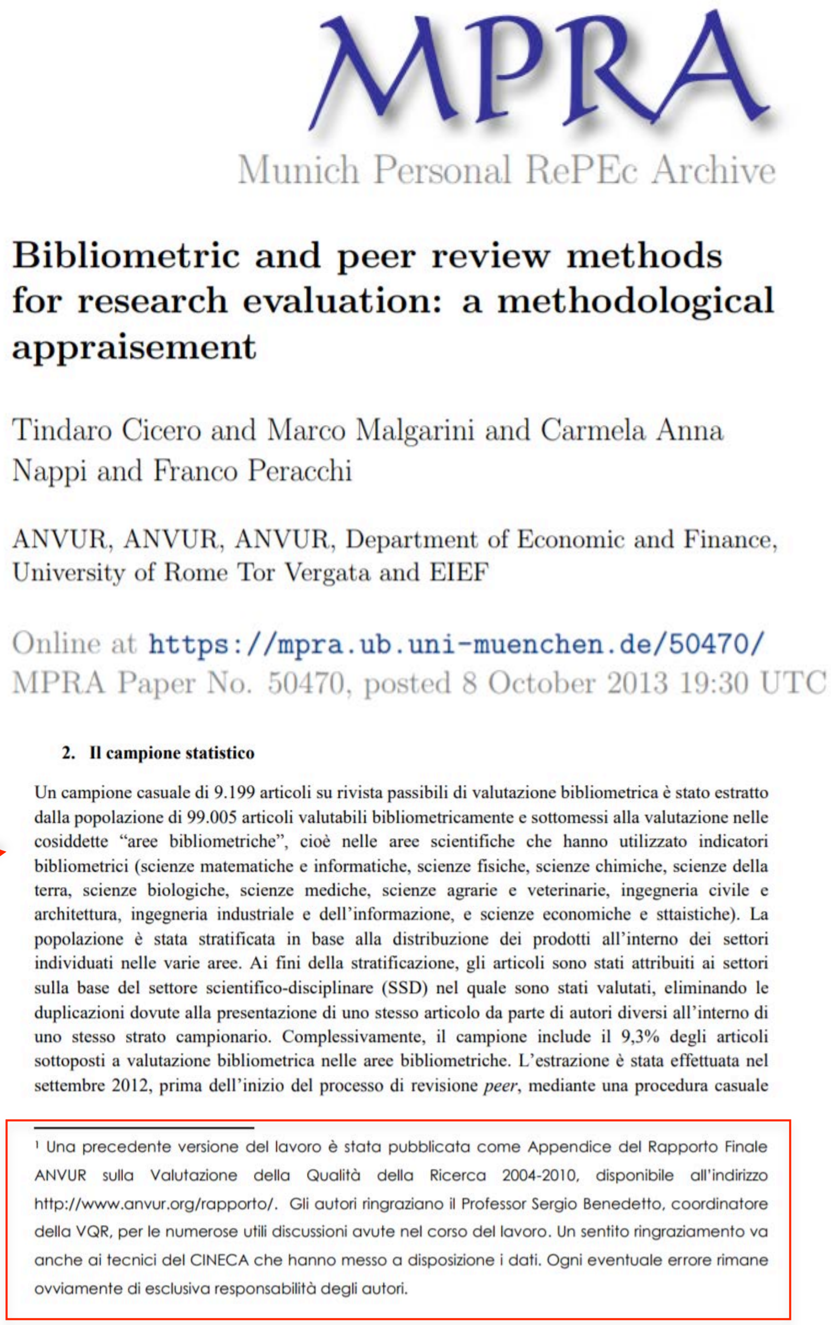
Appendice B. Il confronto tra valutazione *peer* e valutazione bibliometrica

I GEV che hanno utilizzato gli indicatori bibliometrici per la valutazione degli articoli indicizzati in ISI WoS e Scopus hanno selezionato, con un algoritmo di estrazione casuale in grado di garantire una buona copertura statistica di tutti i sub-GEV, un numero pari a circa il 10% degli articoli valutati bibliometricamente e li hanno sottoposti alla valutazione *peer*. L'obiettivo era un confronto tra le due metodologie di valutazione applicate allo stesso campione di articoli, per valutare il grado di corrispondenza dei risultati. Nel seguito, saranno presentati i risultati in forma sintetica e aggregata. Per confronti più puntuali si rimanda alla lettura dell'appendice apposita dei rapporti di area.

B.1 Il campionamento statistico

Un campione casuale di 9199 articoli su rivista passibili di valutazione bibliometrica è stato estratto dalla popolazione di 99005 articoli, valutabili bibliometricamente e sottomessi alla valutazione nei GEV che hanno utilizzato indicatori bibliometrici. La popolazione è stata stratificata in base alla distribuzione dei prodotti all'interno dei sub-GEV individuati nelle varie Aree. Ai fini della stratificazione, gli articoli sono stati attribuiti ai sub-GEV sulla base del settore scientifico-disciplinare (SSD) nel quale sono stati valutati, escludendo i casi di articoli duplicati presentati da diversi autori all'interno di uno stesso strato campionario. Complessivamente, il campione include il 9,3% degli articoli sottoposti a valutazione bibliometrica nelle Aree "bibliometriche". L'estrazione è stata effettuata ai primi di settembre 2012, prima dell'inizio del processo di revisione *peer*, mediante una procedura casuale con il vincolo di selezionare una proporzione significativa di prodotti in ciascun sub-GEV. La Tabella B.1 riporta l'elenco dei GEV bibliometrici e, per ciascuno di essi, la dimensione della popolazione e del campione estratto in valori assoluti e in percentuale sulla popolazione.

1



Bibliometric and peer review methods for research evaluation: a methodological appraisement

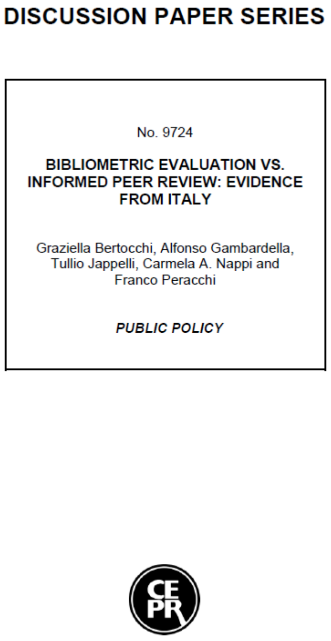
Tindaro Cicero and Marco Malgarini and Carmela Anna Nappi and Franco Peracchi

ANVUR, ANVUR, ANVUR, Department of Economic and Finance, University of Rome Tor Vergata and EIEF

2. Il campione statistico

Un campione casuale di 9.199 articoli su rivista passibili di valutazione bibliometrica è stato estratto dalla popolazione di 99.005 articoli valutabili bibliometricamente e sottomessi alla valutazione nelle cosiddette "aree bibliometriche", cioè nelle aree scientifiche che hanno utilizzato indicatori bibliometrici (scienze matematiche e informatiche, scienze fisiche, scienze chimiche, scienze della terra, scienze biologiche, scienze mediche, scienze agrarie e veterinarie, ingegneria civile e architettura, ingegneria industriale e dell'informazione, e scienze economiche e staistiche). La popolazione è stata stratificata in base alla distribuzione dei prodotti all'interno dei settori individuati nelle varie aree. Ai fini della stratificazione, gli articoli sono stati attribuiti ai settori sulla base del settore scientifico-disciplinare (SSD) nel quale sono stati valutati, eliminando le duplicazioni dovute alla presentazione di uno stesso articolo da parte di autori diversi all'interno di uno stesso strato campionario. Complessivamente, il campione include il 9,3% degli articoli sottoposti a valutazione bibliometrica nelle aree bibliometriche. L'estrazione è stata effettuata nel settembre 2012, prima dell'inizio del processo di revisione *peer*, mediante una procedura casuale

[1] Una precedente versione del lavoro è stata pubblicata come Appendice del Rapporto Finale ANVUR sulla Valutazione della Qualità della Ricerca 2004-2010, disponibile all'indirizzo http://www.anvur.org/rapporto/. Gli autori ringraziano il Professor Sergio Benedetto, coordinatore della VQR, per le numerose utili discussioni avute nel corso del lavoro. Un sentito ringraziamento va anche ai tecnici del CINECA che hanno messo a disposizione i dati. Ogni eventuale errore rimane ovviamente di esclusiva responsabilità degli autori.



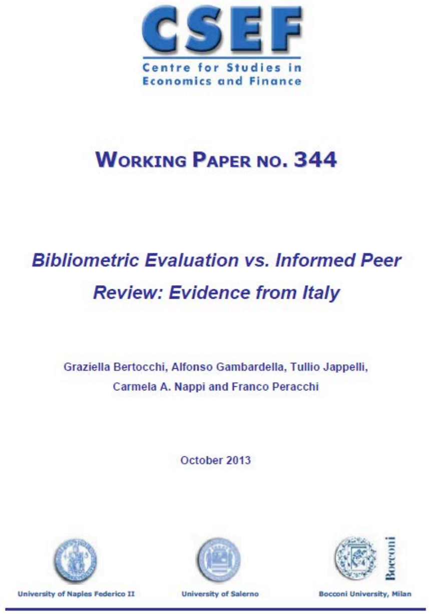
DISCUSSION PAPER SERIES

No. 9724

BIBLIOMETRIC EVALUATION VS. INFORMED PEER REVIEW: EVIDENCE FROM ITALY

Graziella Bertocchi, Alfonso Gambardella, Tullio Jappelli, Carmela A. Nappi and Franco Peracchi

PUBLIC POLICY



CSEF
Centre for Studies in Economics and Finance

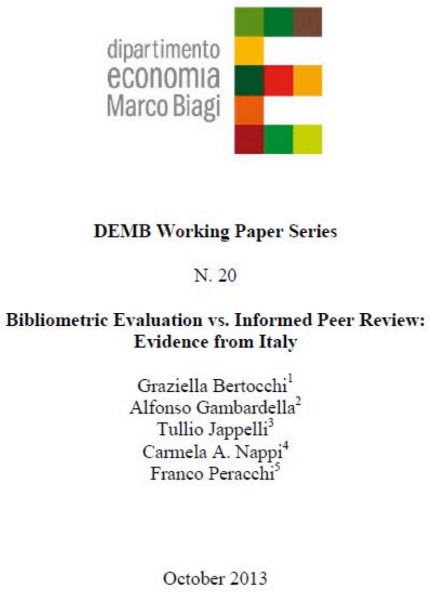
WORKING PAPER NO. 344

Bibliometric Evaluation vs. Informed Peer Review: Evidence from Italy

Graziella Bertocchi, Alfonso Gambardella, Tullio Jappelli, Carmela A. Nappi and Franco Peracchi

October 2013

University of Naples Federico II
University of Salerno
Bocconi University, Milan



dipartimento
economia
Marco Biagi

DEMB Working Paper Series

N. 20

Bibliometric Evaluation vs. Informed Peer Review: Evidence from Italy

Graziella Bertocchi[1]
Alfonso Gambardella[2]
Tullio Jappelli[3]
Carmela A. Nappi[4]
Franco Peracchi[5]

October 2013

DISCUSSION PAPER SERIES

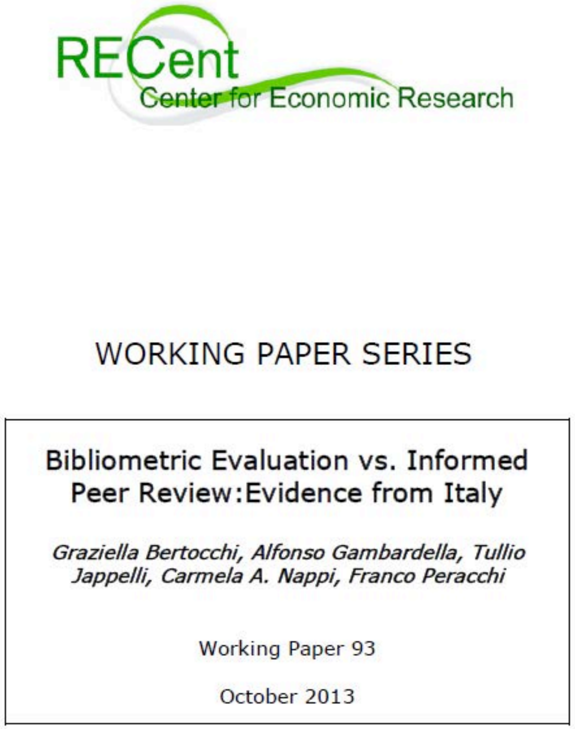
IZA

IZA DP No. 7739

Bibliometric Evaluation vs. Informed Peer Review: Evidence from Italy

Graziella Bertocchi
Alfonso Gambardella
Tullio Jappelli
Carmela A. Nappi
Franco Peracchi

November 2013



RECent
Center for Economic Research

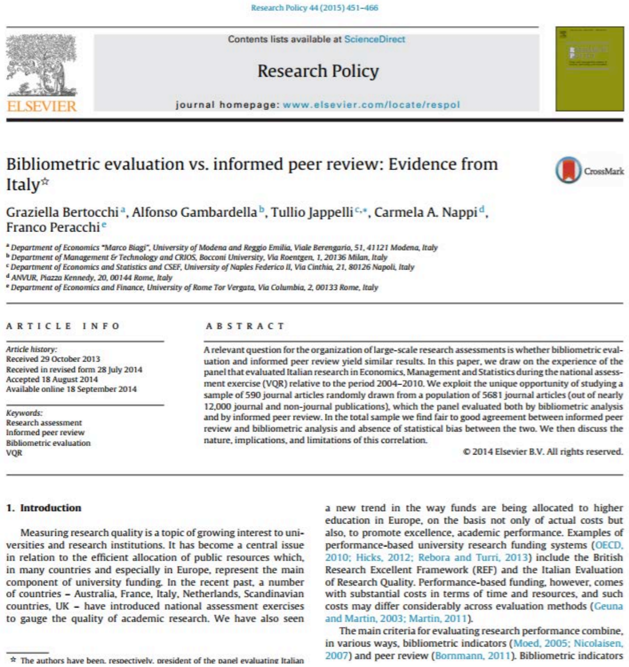
WORKING PAPER SERIES

Bibliometric Evaluation vs. Informed Peer Review:Evidence from Italy

Graziella Bertocchi, Alfonso Gambardella, Tullio Jappelli, Carmela A. Nappi, Franco Peracchi

Working Paper 93

October 2013



Research Policy 44 (2015) 451–466

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

ELSEVIER

Bibliometric evaluation vs. informed peer review: Evidence from Italy☆

Graziella Bertocchi[a], Alfonso Gambardella[b], Tullio Jappelli[c,*], Carmela A. Nappi[d], Franco Peracchi[e]

[a] Department of Economics "Marco Biagi", University of Modena and Reggio Emilia, Viale Berengario, 51, 41121 Modena, Italy
[b] Department of Management & Technology and CRIOS, Bocconi University, Via Roentgen, 1, 20136 Milan, Italy
[c] Department of Economics and Statistics and CSEF, University of Naples Federico II, Via Cinthia, 21, 80126 Napoli, Italy
[d] ANVUR, Piazza Kennedy, 20, 00144 Rome, Italy
[e] Department of Economics and Finance, University of Rome Tor Vergata, Via Columbia, 2, 00133 Rome, Italy

ARTICLE INFO

Article history:
Received 29 October 2013
Received in revised form 28 July 2014
Accepted 18 August 2014
Available online 18 September 2014

Keywords:
Research assessment
Informed peer review
Bibliometric evaluation
VQR

ABSTRACT

A relevant question for the organization of large-scale research assessments is whether bibliometric evaluation and informed peer review yield similar results. In this paper, we draw on the experience of the panel that evaluated Italian research in Economics, Management and Statistics during the national assessment exercise (VQR) relative to the period 2004–2010. We exploit the unique opportunity of studying a sample of 590 journal articles randomly drawn from a population of 5681 journal articles (out of nearly 12,000 journal and non-journal publications), which the panel evaluated both by bibliometric analysis and by informed peer review. In the total sample we find fair to good agreement between informed peer review and bibliometric analysis and absence of statistical bias between the two. We then discuss the nature, implications, and limitations of this correlation.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Measuring research quality is a topic of growing interest to universities and research institutions. It has become a central issue in relation to the efficient allocation of public resources which, in many countries and especially in Europe, represent the main component of university funding. In the recent past, a number of countries – Australia, France, Italy, Netherlands, Scandinavian countries, UK – have introduced national assessment exercises to gauge the quality of academic research. We have also seen a new trend in the way funds are being allocated to higher education in Europe, on the basis not only of actual costs but also, to promote excellence, academic performance. Examples of performance-based university research funding systems (OECD, 2010; Hicks, 2012; Rebora and Turri, 2013) include the British Research Excellent Framework (REF) and the Italian Evaluation of Research Quality. Performance-based funding, however, comes with substantial costs in terms of time and resources, and such costs may differ considerably across evaluation methods (Geuna and Martin, 2003; Martin, 2011).

The main criteria for evaluating research performance combine, in various ways, bibliometric indicators (Moed, 2005; Nicolaisen, 2007) and peer review (Bornmann, 2011). Bibliometric indicators

☆ The authors have been, respectively, president of the panel evaluating Italian

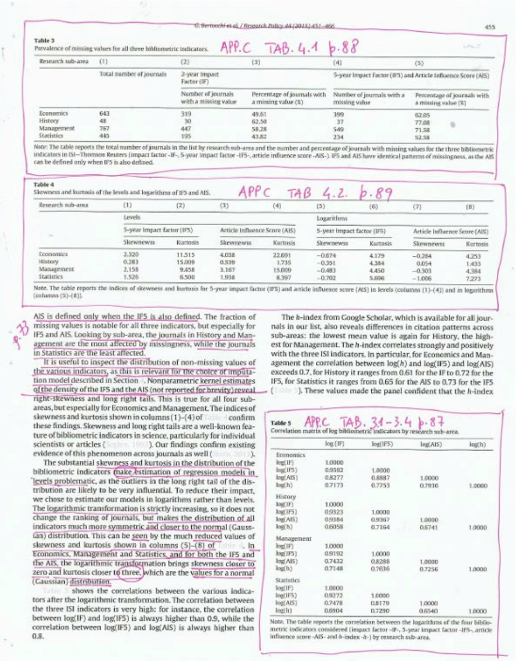


Table 3
Prevalence of missing values for all three bibliometric indicators. APP.C TAB. 4.1 p.88

Research sub-area	(1)	(2)	(3)	(4)	(5)
	Total number of journals	2-year Impact Factor (IF)		5-year Impact Factor (IF5) and Article Influence Score (AIS)	
		Number of journals with a missing value	Percentage of journals with a missing value (%)	Number of journals with a missing value	Percentage of journals with a missing value (%)
Economics	643	319	49.61	399	62.05
History	48	30	62.50	37	77.08
Management	767	447	58.28	549	71.58
Statistics	445	195	43.82	234	52.58

Note: The table reports the total number of journals in the list by research sub-area and the number and percentage of journals with missing values for the three bibliometric indicators in ISI—Thomson Reuters (impact factor -IF-, 5-year impact factor -IF5-, article influence score -AIS-). IF5 and AIS have identical patterns of missingness, as the AIS can be defined only when IF5 is also defined.

Table 4
Skewness and kurtosis of the levels and logarithms of IF5 and AIS. APP C TAB 4.2. p.89

Research sub-area	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Levels				Logarithms			
	5-year Impact factor (IF5)		Article Influence Score (AIS)		5-year Impact factor (IF5)		Article Influence Score (AIS)	
	Skewnewss	Kurtosis	Skewnewss	Kurtosis	Skewnewss	Kurtosis	Skewnewss	Kurtosis
Economics	2.320	11.515	4.038	22.691	−0.674	4.179	−0.284	4.253
History	0.283	15.009	0.539	1.735	−0.351	4.384	0.054	1.433
Management	2.158	9.458	3.167	15.009	−0.483	4.450	−0.303	4.384
Statistics	1.526	6.500	1.938	8.397	−0.702	5.696	−1.006	7.273

Note. The table reports the indices of skewness and kurtosis for 5-year impact factor (IF5) and article influence score (AIS) in levels (columns (1)–(4)) and in logarithms (columns (5)–(8)).

p.73 AIS is defined only when the IF5 is also defined. The fraction of missing values is notable for all three indicators, but especially for IF5 and AIS. Looking by sub-area, the journals in History and Management are the most affected by missingness, while the journals in Statistics are the least affected.

It is useful to inspect the distribution of non-missing values of the various indicators, as this is relevant for the choice of imputation model described in Section 4. Nonparametric kernel estimates of the density of the IF5 and the AIS (not reported for brevity) reveal right-skewness and long right tails. This is true for all four sub-areas, but especially for Economics and Management. The indices of skewness and kurtosis shown in columns (1)–(4) of Table 4 confirm these findings. Skewness and long right tails are a well-known feature of bibliometric indicators in science, particularly for individual scientists or articles (Seglen, 1992). Our findings confirm existing evidence of this phenomenon across journals as well (Stern, 2013).

The substantial skewness and kurtosis in the distribution of the bibliometric indicators make estimation of regression models in levels problematic, as the outliers in the long right tail of the distribution are likely to be very influential. To reduce their impact, we chose to estimate our models in logarithms rather than levels. The logarithmic transformation is strictly increasing, so it does not change the ranking of journals, but makes the distribution of all indicators much more symmetric and closer to the normal (Gaussian) distribution. This can be seen by the much reduced values of skewness and kurtosis shown in columns (5)–(8) of Table 4. In Economics, Management and Statistics, and for both the IF5 and the AIS, the logarithmic transformation brings skewness closer to zero and kurtosis closer to three, which are the values for a normal (Gaussian) distribution.

Table 5 shows the correlations between the various indicators after the logarithmic transformation. The correlation between the three ISI indicators is very high: for instance, the correlation between log(IF) and log(IF5) is always higher than 0.9, while the correlation between log(IF5) and log(AIS) is always higher than 0.8.

The *h*-index from Google Scholar, which is available for all journals in our list, also reveals differences in citation patterns across sub-areas: the lowest mean value is again for History, the highest for Management. The *h*-index correlates strongly and positively with the three ISI indicators. In particular, for Economics and Management the correlation between log(*h*) and log(IF5) and log(AIS) exceeds 0.7, for History it ranges from 0.61 for the IF to 0.72 for the IF5, for Statistics it ranges from 0.65 for the AIS to 0.73 for the IF5 (Table 5). These values made the panel confident that the *h*-index

Table 5 APP.C TAB. 3.1–3.4 p.87
Correlation matrix of log bibliometric indicators by research sub-area.

	log (IF)	log(IF5)	log(AIS)	log(h)
Economics				
log(IF)	1.0000			
log(IF5)	0.9582	1.0000		
log(AIS)	0.8277	0.8887	1.0000	
log(h)	0.7173	0.7753	0.7936	1.0000
History				
log(IF)	1.0000			
log(IF5)	0.9323	1.0000		
log(AIS)	0.9384	0.9367	1.0000	
log(h)	0.6058	0.7164	0.6741	1.0000
Management				
log(IF)	1.0000			
log(IF5)	0.9192	1.0000		
log(AIS)	0.7432	0.8288	1.0000	
log(h)	0.7148	0.7636	0.7256	1.0000
Statistics				
log(IF)	1.0000			
log(IF5)	0.9272	1.0000		
log(AIS)	0.7478	0.8179	1.0000	
log(h)	0.6904	0.7290	0.6540	1.0000

Note. The table reports the correlation between the logarithms of the four bibliometric indicators considered (impact factor -IF-, 5-year impact factor -IF5-, article influence score -AIS- and *h*-index -*h*-) by research sub-area.

APPENDICE C RAPPORTO FINALE AREA 13

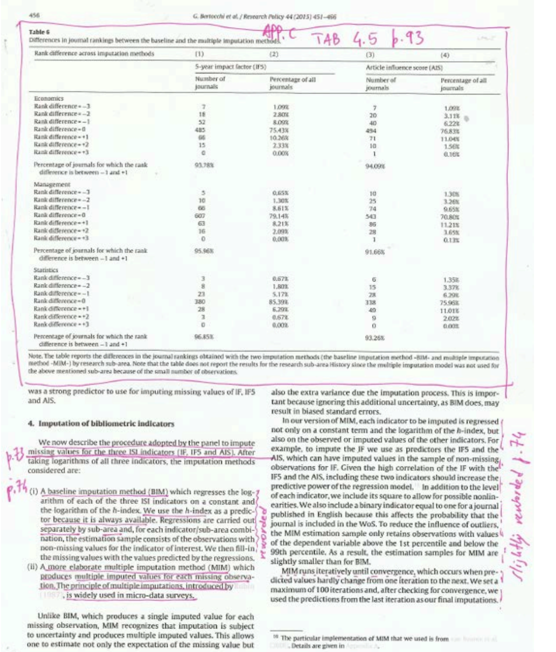


Table 6
Differences in journal rankings between the baseline and the multiple imputation methods. APP.C TAB 4.5 p.93

Rank difference across imputation methods	(1)	(2)	(3)	(4)
	5-year impact factor (IF5)		Article influence score (AIS)	
	Number of journals	Percentage of all journals	Number of journals	Percentage of all journals
Economics				
Rank difference = −3	7	1.09%	7	1.09%
Rank difference = −2	18	2.80%	20	3.11%
Rank difference = −1	52	8.09%	40	6.22%
Rank difference = 0	485	75.43%	494	76.83%
Rank difference = +1	66	10.26%	71	11.04%
Rank difference = +2	15	2.33%	10	1.56%
Rank difference = +3	0	0.00%	1	0.16%
Percentage of journals for which the rank difference is between −1 and +1	93.78%		94.09%	
Management				
Rank difference = −3	5	0.65%	10	1.30%
Rank difference = −2	10	1.30%	25	3.26%
Rank difference = −1	66	8.61%	74	9.65%
Rank difference = 0	607	79.14%	543	70.80%
Rank difference = +1	63	8.21%	86	11.21%
Rank difference = +2	16	2.09%	28	3.65%
Rank difference = +3	0	0.00%	1	0.13%
Percentage of journals for which the rank difference is between −1 and +1	95.96%		91.66%	
Statistics				
Rank difference = −3	3	0.67%	6	1.35%
Rank difference = −2	8	1.80%	15	3.37%
Rank difference = −1	23	5.17%	28	6.29%
Rank difference = 0	380	85.39%	338	75.96%
Rank difference = +1	28	6.29%	49	11.01%
Rank difference = +2	3	0.67%	9	2.02%
Rank difference = +3	0	0.00%	0	0.00%
Percentage of journals for which the rank difference is between −1 and +1	96.85%		93.26%	

Note. The table reports the differences in the journal rankings obtained with the two imputation methods (the baseline imputation method -BIM- and multiple imputation method -MIM-) by research sub-area. Note that the table does not report the results for the research sub-area History since the multiple imputation model was not used for the above mentioned sub-area because of the small number of observations.

was a strong predictor to use for imputing missing values of IF, IF5 and AIS.

4. Imputation of bibliometric indicators

p.73 We now describe the procedure adopted by the panel to impute missing values for the three ISI indicators (IF, IF5 and AIS). After taking logarithms of all three indicators, the imputation methods considered are:

p.74 (i) A baseline imputation method (BIM) which regresses the logarithm of each of the three ISI indicators on a constant and the logarithm of the *h*-index. We use the *h*-index as a predictor because it is always available. Regressions are carried out separately by sub-area and, for each indicator/sub-area combination, the estimation sample consists of the observations with non-missing values for the indicator of interest. We then fill-in the missing values with the values predicted by the regressions. reworded
(ii) A more elaborate multiple imputation method (MIM) which produces multiple imputed values for each missing observation. The principle of multiple imputations, introduced by Rubin (1987), is widely used in micro-data surveys.

Unlike BIM, which produces a single imputed value for each missing observation, MIM recognizes that imputation is subject to uncertainty and produces multiple imputed values. This allows one to estimate not only the expectation of the missing value but also the extra variance due the imputation process. This is important because ignoring this additional uncertainty, as BIM does, may result in biased standard errors.

In our version of MIM, each indicator to be imputed is regressed not only on a constant term and the logarithm of the *h*-index, but also on the observed or imputed values of the other indicators. For example, to impute the IF we use as predictors the IF5 and the AIS, which can have imputed values in the sample of non-missing observations for IF. Given the high correlation of the IF with the IF5 and the AIS, including these two indicators should increase the predictive power of the regression model.[10] In addition to the level of each indicator, we include its square to allow for possible nonlinearities. We also include a binary indicator equal to one for a journal published in English because this affects the probability that the journal is included in the WoS. To reduce the influence of outliers, the MIM estimation sample only retains observations with values of the dependent variable above the 1st percentile and below the 99th percentile. As a result, the estimation samples for MIM are slightly smaller than for BIM. slightly reworded p.74

MIM runs iteratively until convergence, which occurs when predicted values hardly change from one iteration to the next. We set a maximum of 100 iterations and, after checking for convergence, we used the predictions from the last iteration as our final imputations.

[10] The particular implementation of MIM that we used is from van Buuren et al. (2006). Details are given in Appendix A.

APPENDICE A RAPPORTO FINALE

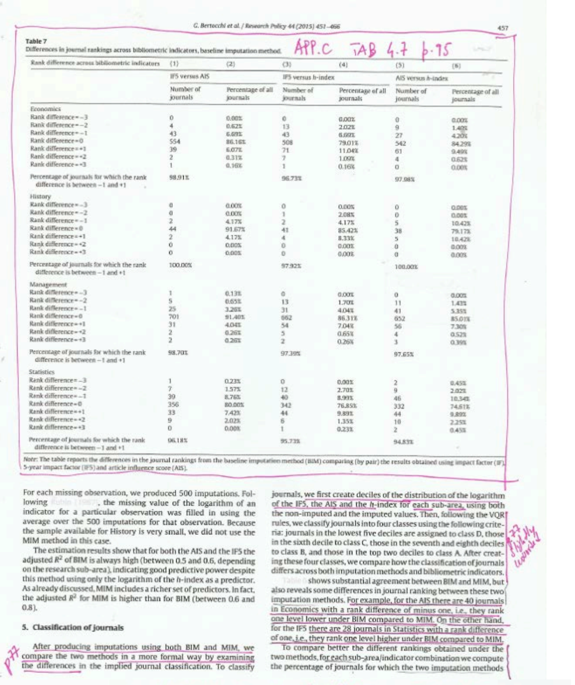


Table 7
Differences in journal rankings across bibliometric indicators, baseline imputation method. APP.C TAB 4.7 p.95

Rank difference across bibliometric indicators	(1)	(2)	(3)	(4)	(5)	(6)
	IF5 versus AIS		IF5 versus *h*-index		AIS versus *h*-index	
	Number of journals	Percentage of all journals	Number of journals	Percentage of all journals	Number of journals	Percentage of all journals
Economics						
Rank difference = −3	0	0.00%	0	0.00%	0	0.00%
Rank difference = −2	4	0.62%	13	2.02%	9	1.40%
Rank difference = −1	43	6.69%	43	6.69%	27	4.20%
Rank difference = 0	554	86.16%	508	79.01%	542	84.29%
Rank difference = +1	39	6.07%	71	11.04%	61	9.49%
Rank difference = +2	2	0.31%	7	1.09%	4	0.62%
Rank difference = +3	1	0.16%	1	0.16%	0	0.00%
Percentage of journals for which the rank difference is between −1 and +1	98.91%		96.73%		97.98%	
History						
Rank difference = −3	0	0.00%	0	0.00%	0	0.00%
Rank difference = −2	0	0.00%	1	2.08%	0	0.00%
Rank difference = −1	2	4.17%	2	4.17%	5	10.42%
Rank difference = 0	44	91.67%	41	85.42%	38	79.17%
Rank difference = +1	2	4.17%	4	8.33%	5	10.42%
Rank difference = +2	0	0.00%	0	0.00%	0	0.00%
Rank difference = +3	0	0.00%	0	0.00%	0	0.00%
Percentage of journals for which the rank difference is between −1 and +1	100.00%		97.92%		100.00%	
Management						
Rank difference = −3	1	0.13%	0	0.00%	0	0.00%
Rank difference = −2	5	0.65%	13	1.70%	11	1.43%
Rank difference = −1	25	3.26%	31	4.04%	41	5.35%
Rank difference = 0	701	91.40%	662	86.31%	652	85.01%
Rank difference = +1	31	4.04%	54	7.04%	56	7.30%
Rank difference = +2	2	0.26%	5	0.65%	4	0.52%
Rank difference = +3	2	0.26%	2	0.26%	3	0.39%
Percentage of journals for which the rank difference is between −1 and +1	98.70%		97.39%		97.65%	
Statistics						
Rank difference = −3	1	0.23%	0	0.00%	2	0.45%
Rank difference = −2	7	1.57%	12	2.70%	9	2.02%
Rank difference = −1	39	8.76%	40	8.99%	46	10.34%
Rank difference = 0	356	80.00%	342	76.85%	332	74.61%
Rank difference = +1	33	7.42%	44	9.89%	44	9.89%
Rank difference = +2	9	2.02%	6	1.35%	10	2.25%
Rank difference = +3	0	0.00%	1	0.23%	2	0.45%
Percentage of journals for which the rank difference is between −1 and +1	96.18%		95.73%		94.83%	

Note: The table reports the differences in the journal rankings from the baseline imputation method (BIM) comparing (by pair) the results obtained using impact factor (IF), 5-year impact factor (IF5) and article influence score (AIS).

For each missing observation, we produced 500 imputations. Following Rubin (1987), the missing value of the logarithm of an indicator for a particular observation was filled in using the average over the 500 imputations for that observation. Because the sample available for History is very small, we did not use the MIM method in this case.

The estimation results show that for both the AIS and the IF5 the adjusted R^2 of BIM is always high (between 0.5 and 0.6, depending on the research sub-area), indicating good predictive power despite this method using only the logarithm of the *h*-index as a predictor. As already discussed, MIM includes a richer set of predictors. In fact, the adjusted R^2 for MIM is higher than for BIM (between 0.6 and 0.8).

5. Classification of journals

p77 After producing imputations using both BIM and MIM, we compare the two methods in a more formal way by examining the differences in the implied journal classification. To classify journals, we first create deciles of the distribution of the logarithm of the IF5, the AIS and the *h*-index for each sub-area, using both the non-imputed and the imputed values. Then, following the VQR rules, we classify journals into four classes using the following criteria: journals in the lowest five deciles are assigned to class D, those in the sixth decile to class C, those in the seventh and eighth deciles to class B, and those in the top two deciles to class A. After creating these four classes, we compare how the classification of journals differs across both imputation methods and bibliometric indicators. p.77 slightly worded

Table 6 shows substantial agreement between BIM and MIM, but also reveals some differences in journal ranking between these two imputation methods. For example, for the AIS there are 40 journals in Economics with a rank difference of minus one, i.e., they rank one level lower under BIM compared to MIM. On the other hand, for the IF5 there are 28 journals in Statistics with a rank difference of one, i.e., they rank one level higher under BIM compared to MIM.

To compare better the different rankings obtained under the two methods, for each sub-area/indicator combination we compute the percentage of journals for which the two imputation methods

APPENDICE A RAPPORTO FINALÉ

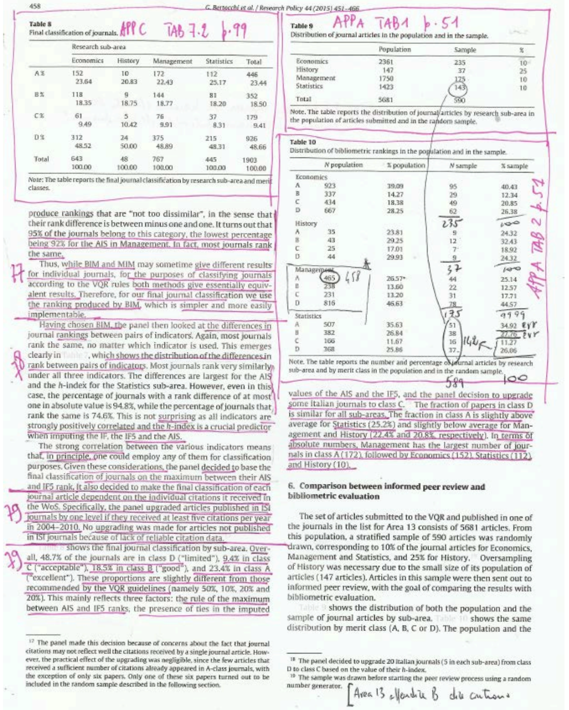


Table 8
Final classification of journals. APP C TAB 7.2 p.99

	Research sub-area				
	Economics	History	Management	Statistics	Total
A %	152 23.64	10 20.83	172 22.43	112 25.17	446 23.44
B %	118 18.35	9 18.75	144 18.77	81 18.20	352 18.50
C %	61 9.49	5 10.42	76 9.91	37 8.31	179 9.41
D %	312 48.52	24 50.00	375 48.89	215 48.31	926 48.66
Total	643 100.00	48 100.00	767 100.00	445 100.00	1903 100.00

Note: The table reports the final journal classification by research sub-area and merit classes.

Table 9 APPA TAB1 p.51
Distribution of journal articles in the population and in the sample.

	Population	Sample	%
Economics	2361	235	10
History	147	37	25
Management	1750	175	10
Statistics	1423	143	10
Total	5681	590	

Note. The table reports the distribution of journal articles by research sub-area in the population of articles submitted and in the random sample.

Table 10 APPA TAB 2 p.52
Distribution of bibliometric rankings in the population and in the sample.

	N population	% population	*N* sample	% sample
Economics				
A	923	39.09	95	40.43
B	337	14.27	29	12.34
C	434	18.38	49	20.85
D	667	28.25	62	26.38
			235	100
History				
A	35	23.81	9	24.32
B	43	29.25	12	32.43
C	25	17.01	7	18.92
D	44	29.93	9	24.32
			37	100
Management				
A	465 458	26.57	44	25.14
B	238	13.60	22	12.57
C	231	13.20	31	17.71
D	816	46.63	78	44.57
			175	99.99
Statistics				
A	507	35.63	51	34.92 err
B	382	26.84	38	27.76 err
C	166	11.67	16	11.27
D	368	25.86	37	26.06

142

Note. The table reports the number and percentage of journal articles by research sub-area and by merit class in the population and in the random sample.

589 100

produce rankings that are "not too dissimilar", in the sense that their rank difference is between minus one and one. It turns out that 95% of the journals belong to this category, the lowest percentage being 92% for the AIS in Management. In fact, most journals rank the same.

77 Thus, while BIM and MIM may sometime give different results for individual journals, for the purposes of classifying journals according to the VQR rules both methods give essentially equivalent results. Therefore, for our final journal classification we use the ranking produced by BIM, which is simpler and more easily implementable.

Having chosen BIM, the panel then looked at the differences in journal rankings between pairs of indicators. Again, most journals rank the same, no matter which indicator is used. This emerges 78 clearly in Table 7, which shows the distribution of the differences in rank between pairs of indicators. Most journals rank very similarly under all three indicators. The differences are largest for the AIS and the *h*-index for the Statistics sub-area. However, even in this case, the percentage of journals with a rank difference of at most one in absolute value is 94.8%, while the percentage of journals that rank the same is 74.6%. This is not surprising as all indicators are strongly positively correlated and the *h*-index is a crucial predictor when imputing the IF, the IF5 and the AIS.

The strong correlation between the various indicators means that, in principle, one could employ any of them for classification purposes. Given these considerations, the panel decided to base the final classification of journals on the maximum between their AIS and IF5 rank. It also decided to make the final classification of each journal article dependent on the individual citations it received in 79 the WoS. Specifically, the panel upgraded articles published in ISI journals by one level if they received at least five citations per year in 2004–2010. No upgrading was made for articles not published in ISI journals because of lack of reliable citation data.[17]

79 Table 8 shows the final journal classification by sub-area. Overall, 48.7% of the journals are in class D ("limited"), 9.4% in class C ("acceptable"), 18.5% in class B ("good"), and 23.4% in class A ("excellent"). These proportions are slightly different from those recommended by the VQR guidelines (namely 50%, 10%, 20% and 20%). This mainly reflects three factors: the rule of the maximum between AIS and IF5 ranks, the presence of ties in the imputed values of the AIS and the IF5, and the panel decision to upgrade some Italian journals to class C.[18] The fraction of papers in class D is similar for all sub-areas. The fraction in class A is slightly above average for Statistics (25.2%) and slightly below average for Management and History (22.4% and 20.8%, respectively). In terms of absolute numbers, Management has the largest number of journals in class A (172), followed by Economics (152), Statistics (112), and History (10).

6. Comparison between informed peer review and bibliometric evaluation

The set of articles submitted to the VQR and published in one of the journals in the list for Area 13 consists of 5681 articles. From this population, a stratified sample of 590 articles was randomly drawn, corresponding to 10% of the journal articles for Economics, Management and Statistics, and 25% for History.[19] Oversampling of History was necessary due to the small size of its population of articles (147 articles). Articles in this sample were then sent out to informed peer review, with the goal of comparing the results with bibliometric evaluation.

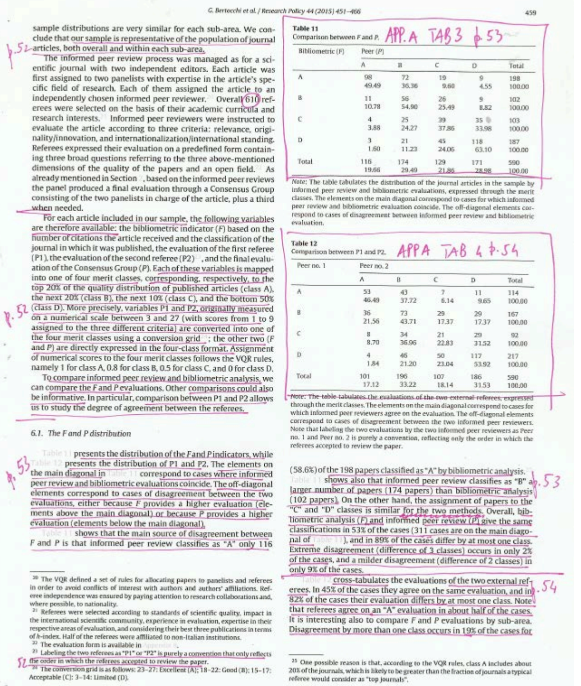
Table 9 shows the distribution of both the population and the sample of journal articles by sub-area. Table 10 shows the same distribution by merit class (A, B, C or D). The population and the

[17] The panel made this decision because of concerns about the fact that journal citations may not reflect well the citations received by a single journal article. However, the practical effect of the upgrading was negligible, since the few articles that received a sufficient number of citations already appeared in A-class journals, with the exception of only six papers. Only one of these six papers turned out to be included in the random sample described in the following section.

[18] The panel decided to upgrade 20 Italian journals (5 in each sub-area) from class D to class C based on the value of their *h*-index.

[19] The sample was drawn before starting the peer review process using a random number generator. [Area 13 allegato B dei criteri

APPENDICE C RAPPORTO FINALE



sample distributions are very similar for each sub-area. We conclude that our sample is representative of the population of journal p.52 articles, both overall and within each sub-area.

The informed peer review process was managed as for a scientific journal with two independent editors. Each article was first assigned to two panelists with expertise in the article's specific field of research. Each of them assigned the article to an independently chosen informed peer reviewer.[20] Overall 610 referees were selected on the basis of their academic curricula and research interests.[21] Informed peer reviewers were instructed to evaluate the article according to three criteria: relevance, originality/innovation, and internationalization/international standing. Referees expressed their evaluation on a predefined form containing three broad questions referring to the three above-mentioned dimensions of the quality of the papers and an open field.[22] As already mentioned in Section 2, based on the informed peer reviews the panel produced a final evaluation through a Consensus Group consisting of the two panelists in charge of the article, plus a third when needed.

For each article included in our sample, the following variables are therefore available: the bibliometric indicator (*F*) based on the number of citations the article received and the classification of the journal in which it was published, the evaluation of the first referee (P1), the evaluation of the second referee (P2)[23], and the final evaluation of the Consensus Group (*P*). Each of these variables is mapped into one of four merit classes, corresponding, respectively, to the top 20% of the quality distribution of published articles (class A), the next 20% (class B), the next 10% (class C), and the bottom 50% p.52 (class D). More precisely, variables P1 and P2, originally measured on a numerical scale between 3 and 27 (with scores from 1 to 9 assigned to the three different criteria) are converted into one of the four merit classes using a conversion grid[24]; the other two (*F* and *P*) are directly expressed in the four-class format. Assignment of numerical scores to the four merit classes follows the VQR rules, namely 1 for class A, 0.8 for class B, 0.5 for class C, and 0 for class D.

To compare informed peer review and bibliometric analysis, we can compare the *F* and *P* evaluations. Other comparisons could also be informative. In particular, comparison between P1 and P2 allows us to study the degree of agreement between the referees.

6.1. The F and P distribution

Table 11 presents the distribution of the *F* and *P* indicators, while p.53 Table 12 presents the distribution of P1 and P2. The elements on the main diagonal in Table 11 correspond to cases where informed peer review and bibliometric evaluations coincide. The off-diagonal elements correspond to cases of disagreement between the two evaluations, either because *F* provides a higher evaluation (elements above the main diagonal) or because *P* provides a higher evaluation (elements below the main diagonal).

Table 11 shows that the main source of disagreement between *F* and *P* is that informed peer review classifies as "A" only 116 (58.6%) of the 198 papers classified as "A" by bibliometric analysis.[25] Table 11 shows also that informed peer review classifies as "B" a larger number of papers (174 papers) than bibliometric analysis (102 papers). On the other hand, the assignment of papers to the "C" and "D" classes is similar for the two methods. Overall, bibliometric analysis (*F*) and informed peer review (*P*) give the same classifications in 53% of the cases (311 cases are on the main diagonal of Table 11), and in 89% of the cases differ by at most one class. Extreme disagreement (difference of 3 classes) occurs in only 2% of the cases, and a milder disagreement (difference of 2 classes) in only 9% of the cases. p.53

Table 12 cross-tabulates the evaluations of the two external referees. In 45% of the cases they agree on the same evaluation, and in 82% of the cases their evaluation differs by at most one class. Note that referees agree on an "A" evaluation in about half of the cases. p.54 It is interesting also to compare *F* and *P* evaluations by sub-area. Disagreement by more than one class occurs in 19% of the cases for

Table 11
Comparison between *F* and *P*. APP.A TAB3 p.53

Bibliometric (*F*)	Peer (*P*)				
	A	B	C	D	Total
A	98 49.49	72 36.36	19 9.60	9 4.55	198 100.00
B	11 10.78	56 54.90	26 25.49	9 8.82	102 100.00
C	4 3.88	25 24.27	39 37.86	35 33.98	103 100.00
D	3 1.60	21 11.23	45 24.06	118 63.10	187 100.00
Total	116 19.66	174 29.49	129 21.86	171 28.98	590 100.00

Note: The table tabulates the distribution of the journal articles in the sample by informed peer review and bibliometric evaluations, expressed through the merit classes. The elements on the main diagonal correspond to cases for which informed peer review and bibliometric evaluation coincide. The off-diagonal elements correspond to cases of disagreement between informed peer review and bibliometric evaluation.

Table 12
Comparison between P1 and P2. APPA TAB 4 p.54

Peer no. 1	Peer no. 2				
	A	B	C	D	Total
A	53 46.49	43 37.72	7 6.14	11 9.65	114 100.00
B	36 21.56	73 43.71	29 17.37	29 17.37	167 100.00
C	8 8.70	34 36.96	21 22.83	29 31.52	92 100.00
D	4 1.84	46 21.20	50 23.04	117 53.92	217 100.00
Total	101 17.12	196 33.22	107 18.14	186 31.53	590 100.00

~~*Note*: The table tabulates the evaluations of the two external referees, expressed~~ through the merit classes. The elements on the main diagonal correspond to cases for which informed peer reviewers agree on the evaluation. The off-diagonal elements correspond to cases of disagreement between the two informed peer reviewers. Note that labeling the two evaluations by the two informed peer reviewers as Peer no. 1 and Peer no. 2 is purely a convention, reflecting only the order in which the referees accepted to review the paper.

[20] The VQR defined a set of rules for allocating papers to panelists and referees in order to avoid conflicts of interest with authors and authors' affiliations. Referee independence was ensured by paying attention to research collaborations and, where possible, to nationality.

[21] Referees were selected according to standards of scientific quality, impact in the international scientific community, experience in evaluation, expertise in their respective areas of evaluation, and considering their best three publications in terms of *h*-index. Half of the referees were affiliated to non-Italian institutions.

[22] The evaluation form is available in Appendix B.

[23] Labeling the two referees as "P1" or "P2" is purely a convention that only reflects 52 the order in which the referees accepted to review the paper.

[24] The conversion grid is as follows: 23–27: Excellent (A); 18–22: Good (B); 15–17: Acceptable (C); 3–14: Limited (D).

[25] One possible reason is that, according to the VQR rules, class A includes about 20% of the journals, which is likely to be greater than the fraction of journals a typical referee would consider as "top journals".

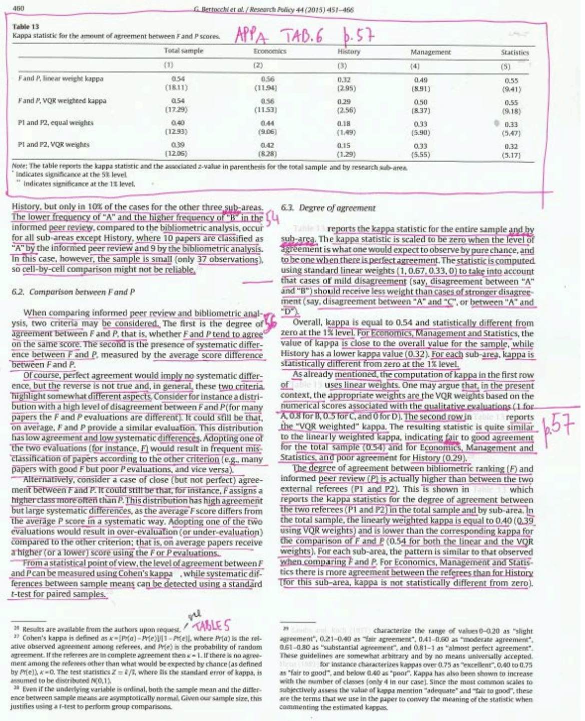


Table 13
Kappa statistic for the amount of agreement between *F* and *P* scores. APPA TAB.6 p.57

	Total sample	Economics	History	Management	Statistics
	(1)	(2)	(3)	(4)	(5)
F and *P*, linear weight kappa	0.54 (18.11)	0.56 (11.94)	0.32 (2.95)	0.49 (8.91)	0.55 (9.41)
F and *P*, VQR weighted kappa	0.54 (17.29)	0.56 (11.53)	0.29 (2.56)	0.50 (8.37)	0.55 (9.18)
P1 and P2, equal weights	0.40 (12.93)	0.44 (9.06)	0.18 (1.49)	0.33 (5.90)	0.33 (5.47)
P1 and P2, VQR weights	0.39 (12.06)	0.42 (8.28)	0.15 (1.29)	0.33 (5.55)	0.32 (5.17)

Note: The table reports the kappa statistic and the associated *z*-value in parenthesis for the total sample and by research sub-area.
* Indicates significance at the 5% level.
** Indicates significance at the 1% level.

History, but only in 10% of the cases for the other three sub-areas. The lower frequency of "A" and the higher frequency of "B" in the 54 informed peer review, compared to the bibliometric analysis, occur for all sub-areas except History, where 10 papers are classified as "A" by the informed peer review and 9 by the bibliometric analysis. In this case, however, the sample is small (only 37 observations), so cell-by-cell comparison might not be reliable.[26]

6.2. Comparison between F and P

When comparing informed peer review and bibliometric analysis, two criteria may be considered. The first is the degree of 56 agreement between *F* and *P*, that is, whether *F* and *P* tend to agree on the same score. The second is the presence of systematic difference between *F* and *P*, measured by the average score difference between *F* and *P*.

Of course, perfect agreement would imply no systematic difference, but the reverse is not true and, in general, these two criteria highlight somewhat different aspects. Consider for instance a distribution with a high level of disagreement between *F* and *P* (for many papers the *F* and *P* evaluations are different). It could still be that, on average, *F* and *P* provide a similar evaluation. This distribution has low agreement and low systematic differences. Adopting one of the two evaluations (for instance, *F*) would result in frequent misclassification of papers according to the other criterion (e.g., many papers with good *F* but poor *P* evaluations, and vice versa).

Alternatively, consider a case of close (but not perfect) agreement between *F* and *P*. It could still be that, for instance, *F* assigns a higher class more often than *P*. This distribution has high agreement but large systematic differences, as the average *F* score differs from the average *P* score in a systematic way. Adopting one of the two evaluations would result in over-evaluation (or under-evaluation) compared to the other criterion; that is, on average papers receive a higher (or a lower) score using the *F* or *P* evaluations.

From a statistical point of view, the level of agreement between *F* and *P* can be measured using Cohen's kappa[27], while systematic differences between sample means can be detected using a standard *t*-test for paired samples.[28]

6.3. Degree of agreement

Table 13 reports the kappa statistic for the entire sample and by sub-area. The kappa statistic is scaled to be zero when the level of agreement is what one would expect to observe by pure chance, and to be one when there is perfect agreement. The statistic is computed using standard linear weights (1, 0.67, 0.33, 0) to take into account that cases of mild disagreement (say, disagreement between "A" and "B") should receive less weight than cases of stronger disagreement (say, disagreement between "A" and "C", or between "A" and "D").

Overall, kappa is equal to 0.54 and statistically different from zero at the 1% level. For Economics, Management and Statistics, the value of kappa is close to the overall value for the sample, while History has a lower kappa value (0.32). For each sub-area, kappa is statistically different from zero at the 1% level.

As already mentioned, the computation of kappa in the first row of Table 13 uses linear weights. One may argue that, in the present context, the appropriate weights are the VQR weights based on the numerical scores associated with the qualitative evaluations (1 for A, 0.8 for B, 0.5 for C, and 0 for D). The second row in Table 13 reports the "VQR weighted" kappa. The resulting statistic is quite similar to the linearly weighted kappa, indicating fair to good agreement for the total sample (0.54) and for Economics, Management and Statistics, and poor agreement for History (0.29).[29] p.57

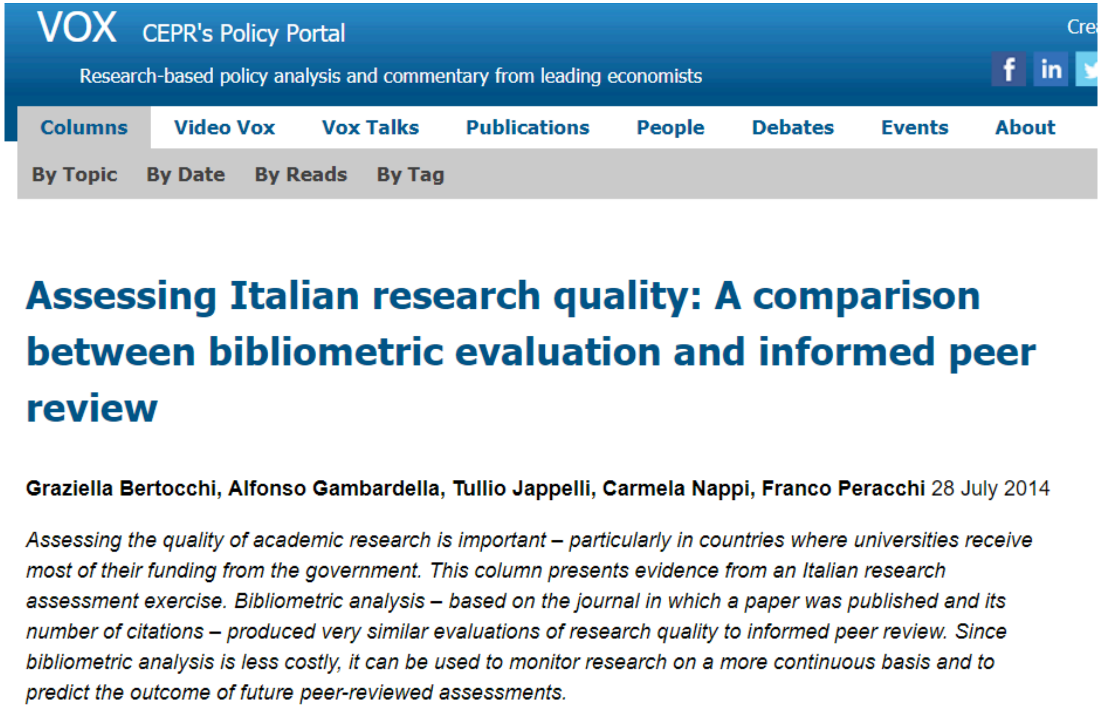
The degree of agreement between bibliometric ranking (*F*) and informed peer review (*P*) is actually higher than between the two external referees (P1 and P2). This is shown in Table 13 which reports the kappa statistics for the degree of agreement between the two referees (P1 and P2) in the total sample and by sub-area. In the total sample, the linearly weighted kappa is equal to 0.40 (0.39 using VQR weights) and is lower than the corresponding kappa for the comparison of *F* and *P* (0.54 for both the linear and the VQR weights). For each sub-area, the pattern is similar to that observed when comparing *F* and *P*. For Economics, Management and Statistics there is more agreement between the referees than for History (for this sub-area, kappa is not statistically different from zero).

[26] Results are available from the authors upon request. / TABLE 5

[27] Cohen's kappa is defined as $\kappa = [Pr(a) - Pr(e)]/[1 - Pr(e)]$, where $Pr(a)$ is the relative observed agreement among referees, and $Pr(e)$ is the probability of random agreement. If the referees are in complete agreement then $\kappa = 1$. If there is no agreement among the referees other than what would be expected by chance (as defined by $Pr(e)$), $\kappa = 0$. The test statistics $Z = \hat{\kappa}/\hat{s}$, where $\hat{s}$ is the standard error of kappa, is assumed to be distributed $N(0,1)$.

[28] Even if the underlying variable is ordinal, both the sample mean and the difference between sample means are asymptotically normal. Given our sample size, this justifies using a *t*-test to perform group comparisons.

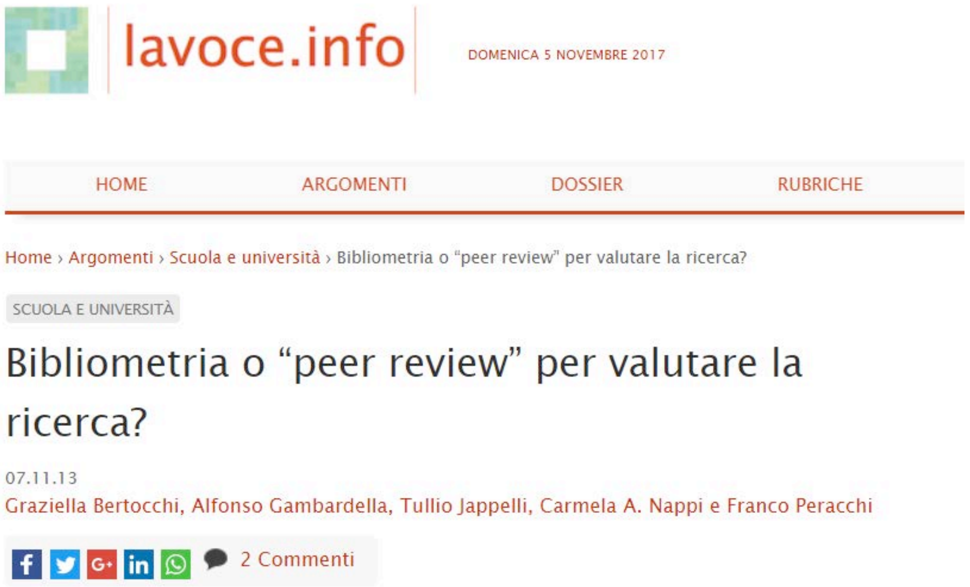
[29] Landis and Koch (1977) characterize the range of values 0–0.20 as "slight agreement", 0.21–0.40 as "fair agreement", 0.41–0.60 as "moderate agreement", 0.61–0.80 as "substantial agreement", and 0.81–1 as "almost perfect agreement". These guidelines are somewhat arbitrary and by no means universally accepted. Fleiss (1981) for instance characterizes kappas over 0.75 as "excellent", 0.40 to 0.75 as "fair to good", and below 0.40 as "poor". Kappa has also been shown to increase with the number of classes (only 4 in our case). Since the most common scales to subjectively assess the value of kappa mention "adequate" and "fair to good", these are the terms that we use in the paper to convey the meaning of the statistic when commenting the estimated kappas.



Assessing Italian research quality: A comparison between bibliometric evaluation and informed peer review

Graziella Bertocchi, Alfonso Gambardella, Tullio Jappelli, Carmela Nappi, Franco Peracchi 28 July 2014

Assessing the quality of academic research is important – particularly in countries where universities receive most of their funding from the government. This column presents evidence from an Italian research assessment exercise. Bibliometric analysis – based on the journal in which a paper was published and its number of citations – produced very similar evaluations of research quality to informed peer review. Since bibliometric analysis is less costly, it can be used to monitor research on a more continuous basis and to predict the outcome of future peer-reviewed assessments.



Home › Argomenti › Scuola e università › Bibliometria o "peer review" per valutare la ricerca?

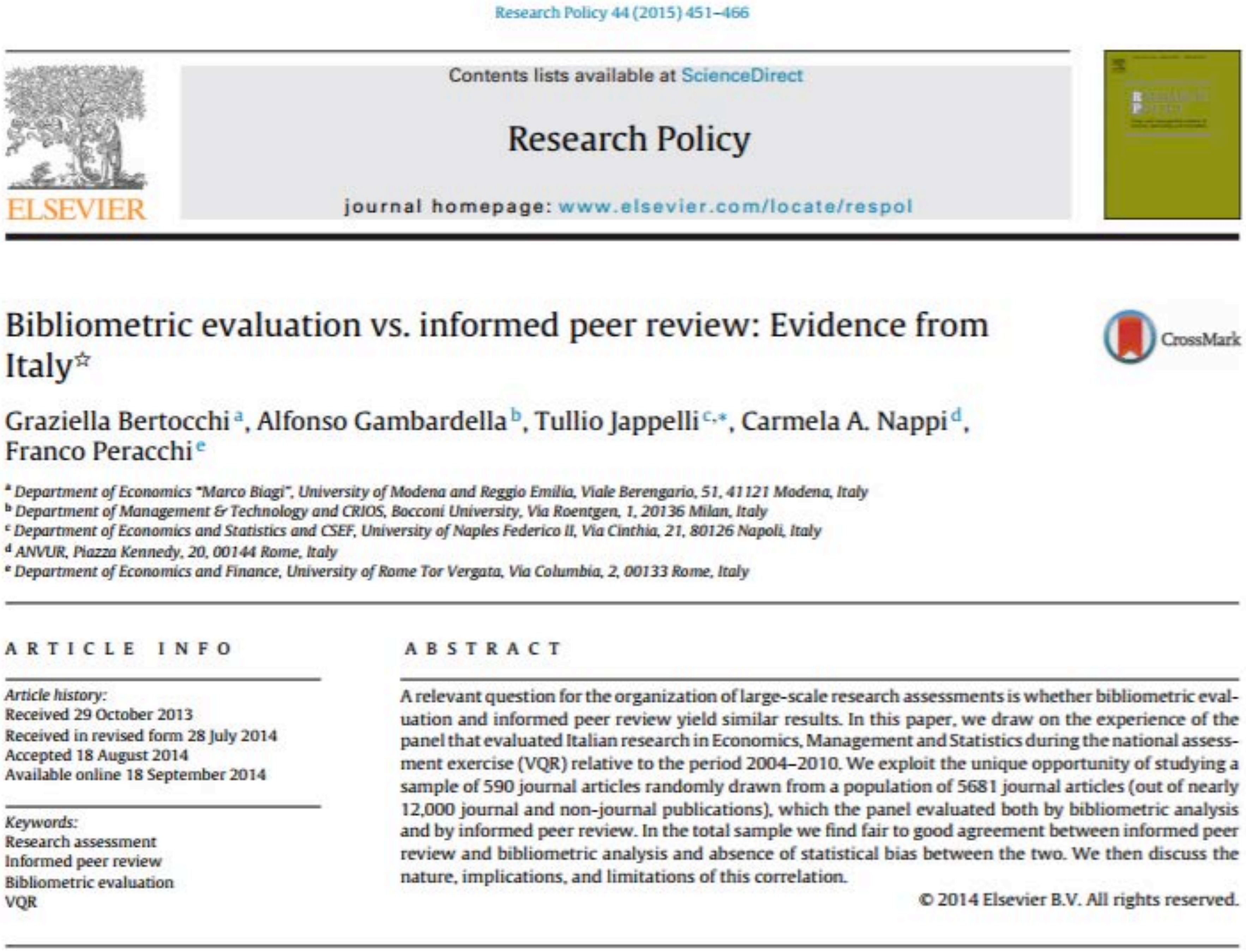
SCUOLA E UNIVERSITÀ

Bibliometria o "peer review" per valutare la ricerca?

07.11.13

Graziella Bertocchi, Alfonso Gambardella, Tullio Jappelli, Carmela A. Nappi e Franco Peracchi

4. Bibliometrics vs peer review: do they agree?



Research Policy 44 (2015) 451–466

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Bibliometric evaluation vs. informed peer review: Evidence from Italy☆

Graziella Bertocchi[a], Alfonso Gambardella[b], Tullio Jappelli[c,*], Carmela A. Nappi[d], Franco Peracchi[e]

[a] Department of Economics "Marco Biagi", University of Modena and Reggio Emilia, Viale Berengario, 51, 41121 Modena, Italy
[b] Department of Management & Technology and CRIOS, Bocconi University, Via Roentgen, 1, 20136 Milan, Italy
[c] Department of Economics and Statistics and CSEF, University of Naples Federico II, Via Cinthia, 21, 80126 Napoli, Italy
[d] ANVUR, Piazza Kennedy, 20, 00144 Rome, Italy
[e] Department of Economics and Finance, University of Rome Tor Vergata, Via Columbia, 2, 00133 Rome, Italy

ARTICLE INFO

Article history:
Received 29 October 2013
Received in revised form 28 July 2014
Accepted 18 August 2014
Available online 18 September 2014

Keywords:
Research assessment
Informed peer review
Bibliometric evaluation
VQR

ABSTRACT

A relevant question for the organization of large-scale research assessments is whether bibliometric evaluation and informed peer review yield similar results. In this paper, we draw on the experience of the panel that evaluated Italian research in Economics, Management and Statistics during the national assessment exercise (VQR) relative to the period 2004–2010. We exploit the unique opportunity of studying a sample of 590 journal articles randomly drawn from a population of 5681 journal articles (out of nearly 12,000 journal and non-journal publications), which the panel evaluated both by bibliometric analysis and by informed peer review. In the total sample we find fair to good agreement between informed peer review and bibliometric analysis and absence of statistical bias between the two. We then discuss the nature, implications, and limitations of this correlation.

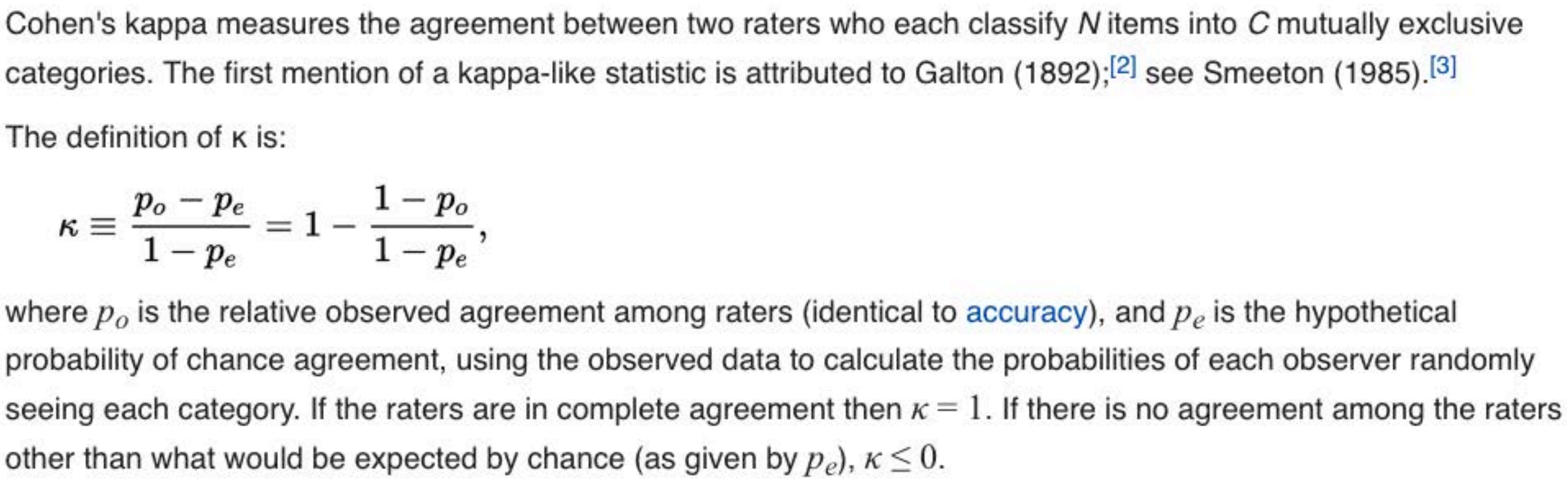
Table 11

Comparison between *F* and *P*.

Bibliometric (*F*)	Peer (*P*)				
	A	B	C	D	Total
A	98 49.49	72 36.36	19 9.60	9 4.55	198 100.00
B	11 10.78	56 54.90	26 25.49	9 8.82	102 100.00
C	4 3.88	25 24.27	39 37.86	35 33.98	103 100.00
D	3 1.60	21 11.23	45 24.06	118 63.10	187 100.00
Total	116 19.66	174 29.49	129 21.86	171 28.98	590 100.00

Note: The table tabulates the distribution of the journal articles in the sample by informed peer review and bibliometric evaluations, expressed through the merit classes. The elements on the main diagonal correspond to cases for which informed peer review and bibliometric evaluation coincide. The off-diagonal elements correspond to cases of disagreement between informed peer review and bibliometric evaluation.

Cohen’s kappa



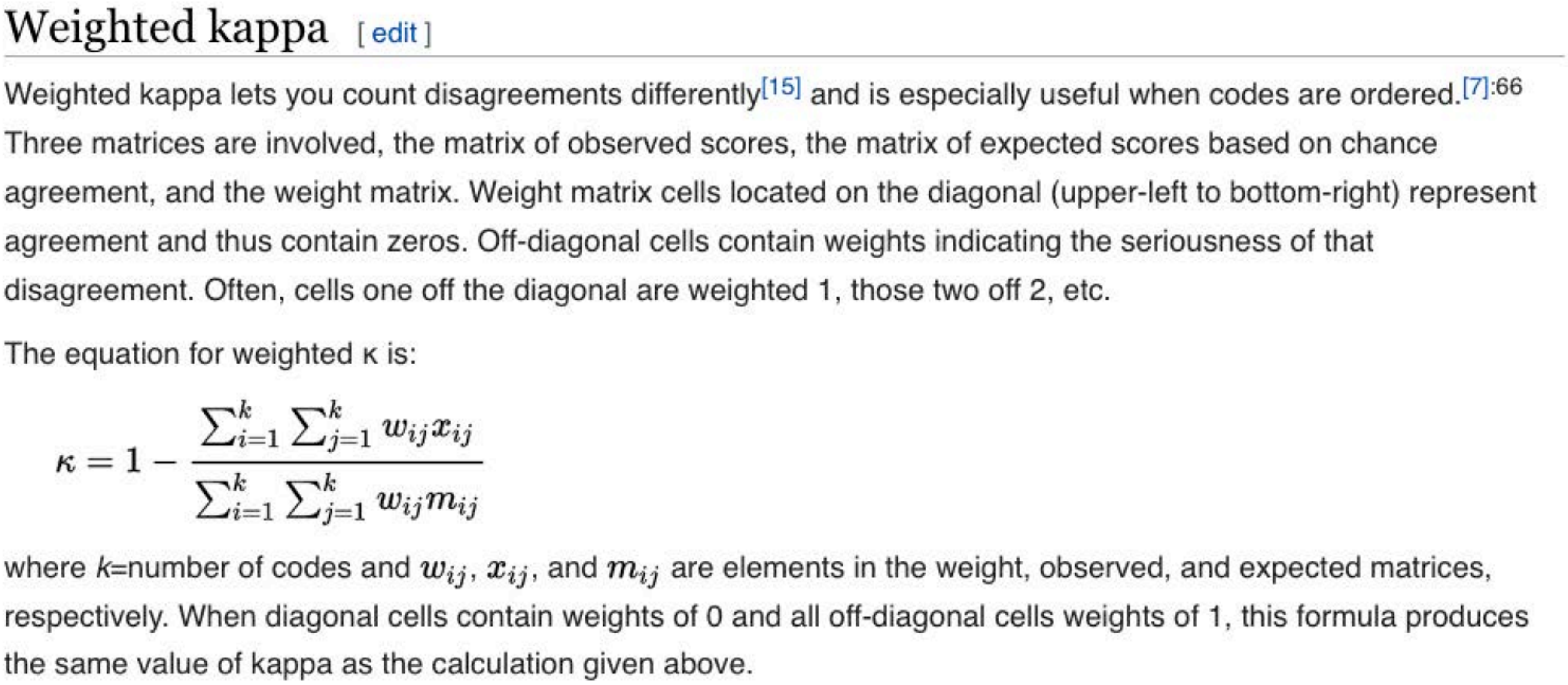
Cohen's kappa measures the agreement between two raters who each classify *N* items into *C* mutually exclusive categories. The first mention of a kappa-like statistic is attributed to Galton (1892);[2] see Smeeton (1985).[3]

The definition of κ is:

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

where p_o is the relative observed agreement among raters (identical to accuracy), and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters other than what would be expected by chance (as given by p_e), $\kappa \leq 0$.

Weighted Cohen’s kappa



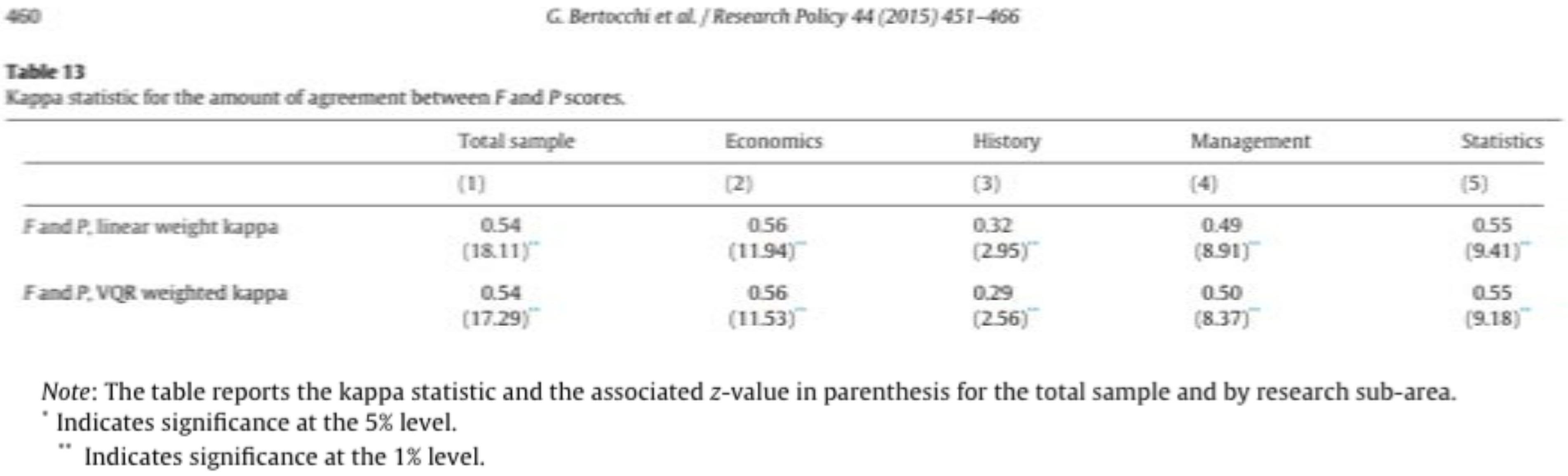
Weighted kappa [edit]

Weighted kappa lets you count disagreements differently[15] and is especially useful when codes are ordered.[7]:66 Three matrices are involved, the matrix of observed scores, the matrix of expected scores based on chance agreement, and the weight matrix. Weight matrix cells located on the diagonal (upper-left to bottom-right) represent agreement and thus contain zeros. Off-diagonal cells contain weights indicating the seriousness of that disagreement. Often, cells one off the diagonal are weighted 1, those two off 2, etc.

The equation for weighted κ is:

$$\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}}$$

where *k*=number of codes and w_{ij}, x_{ij}, and m_{ij} are elements in the weight, observed, and expected matrices, respectively. When diagonal cells contain weights of 0 and all off-diagonal cells weights of 1, this formula produces the same value of kappa as the calculation given above.



460 G. Bertocchi et al. / Research Policy 44 (2015) 451–466

Table 13
Kappa statistic for the amount of agreement between F and P scores.

	Total sample	Economics	History	Management	Statistics
	(1)	(2)	(3)	(4)	(5)
F and P, linear weight kappa	0.54 (18.11)**	0.56 (11.94)**	0.32 (2.95)**	0.49 (8.91)**	0.55 (9.41)**
F and P, VQR weighted kappa	0.54 (17.29)**	0.56 (11.53)**	0.29 (2.56)**	0.50 (8.37)**	0.55 (9.18)**

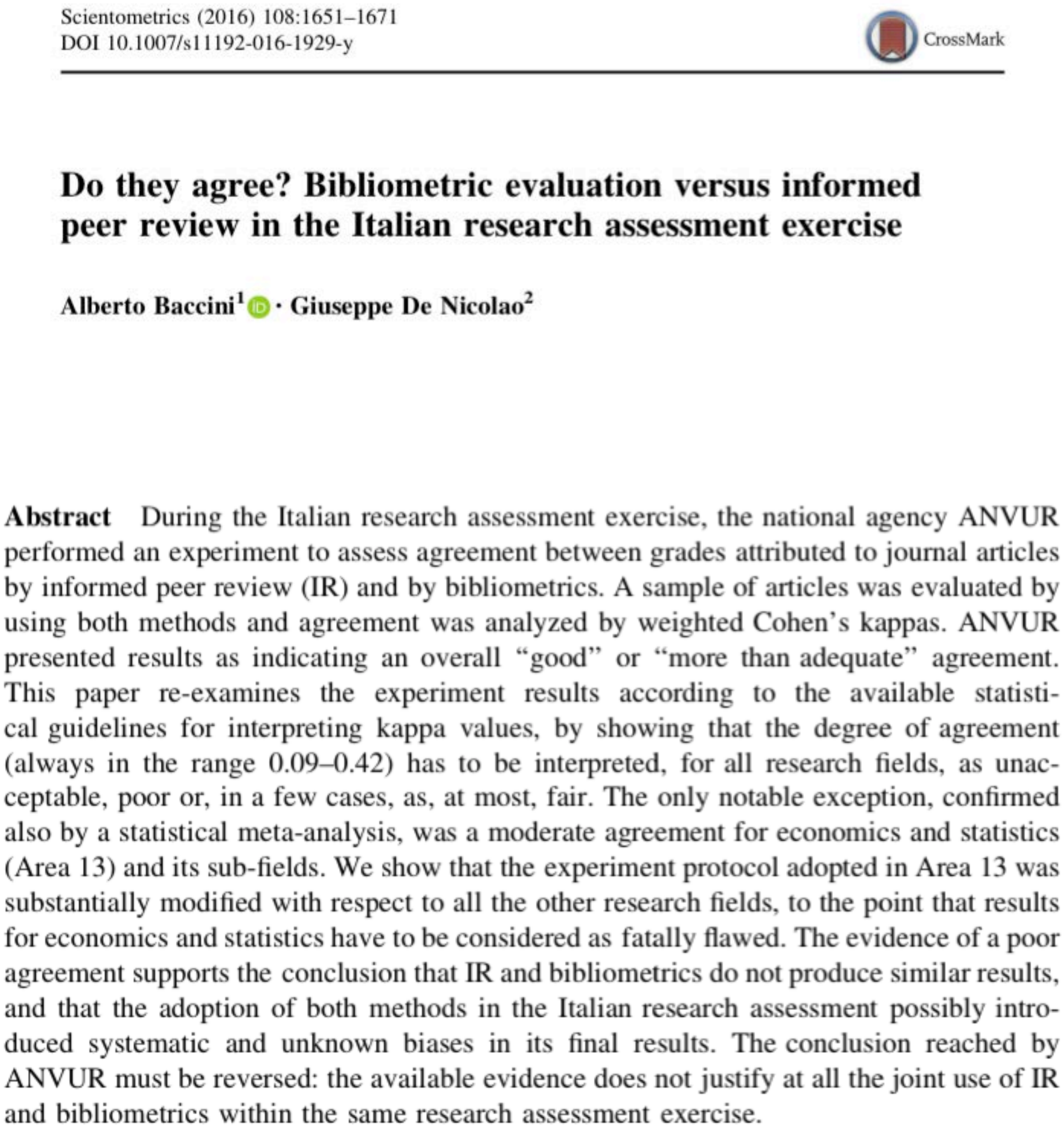
Note: The table reports the kappa statistic and the associated *z*-value in parenthesis for the total sample and by research sub-area.
* Indicates significance at the 5% level.
** Indicates significance at the 1% level.

«*The second row in Table 13 reports the "VQR weighted" kappa. The resulting statistic is quite similar to the linearly weighted kappa, indicating* ***fair to good agreement*** *for the total sample (****0.54****) and for Economics, Management and Statistics, and poor agreement for History (0.29)*.»

Therefore:

“the agencies that run these evaluations could feel confident about using bibliometric evaluations and interpret the results as highly correlated with what they would obtain if they performed informed peer review” (Bertocchi et al. 2015)

Is this true?



Scientometrics (2016) 108:1651–1671
DOI 10.1007/s11192-016-1929-y

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

Alberto Baccini[1] · Giuseppe De Nicolao[2]

Abstract During the Italian research assessment exercise, the national agency ANVUR performed an experiment to assess agreement between grades attributed to journal articles by informed peer review (IR) and by bibliometrics. A sample of articles was evaluated by using both methods and agreement was analyzed by weighted Cohen's kappas. ANVUR presented results as indicating an overall "good" or "more than adequate" agreement. This paper re-examines the experiment results according to the available statistical guidelines for interpreting kappa values, by showing that the degree of agreement (always in the range 0.09–0.42) has to be interpreted, for all research fields, as unacceptable, poor or, in a few cases, as, at most, fair. The only notable exception, confirmed also by a statistical meta-analysis, was a moderate agreement for economics and statistics (Area 13) and its sub-fields. We show that the experiment protocol adopted in Area 13 was substantially modified with respect to all the other research fields, to the point that results for economics and statistics have to be considered as fatally flawed. The evidence of a poor agreement supports the conclusion that IR and bibliometrics do not produce similar results, and that the adoption of both methods in the Italian research assessment possibly introduced systematic and unknown biases in its final results. The conclusion reached by ANVUR must be reversed: the available evidence does not justify at all the joint use of IR and bibliometrics within the same research assessment exercise.

Concordanza: “fair to good”. Ma quanto “good”?

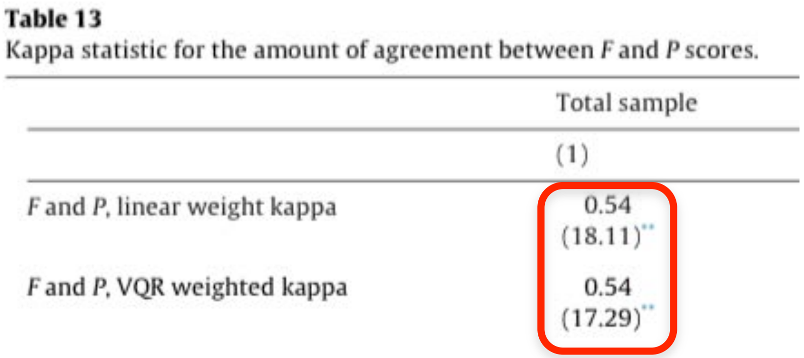
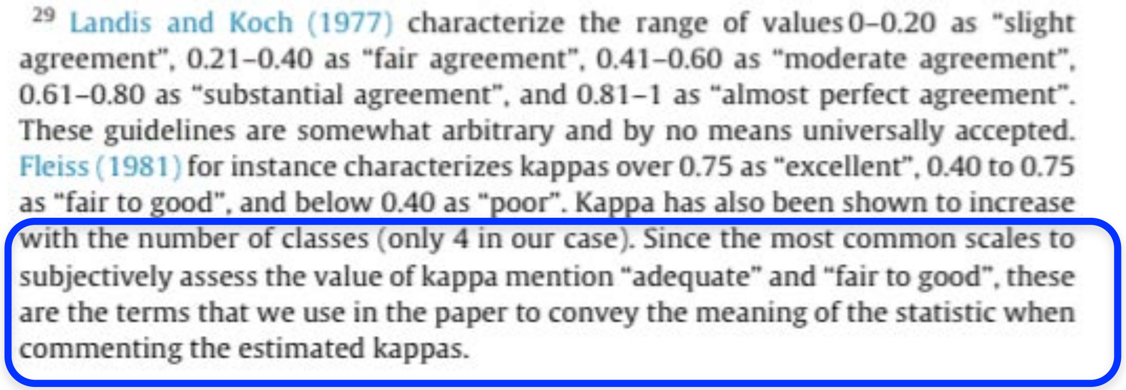
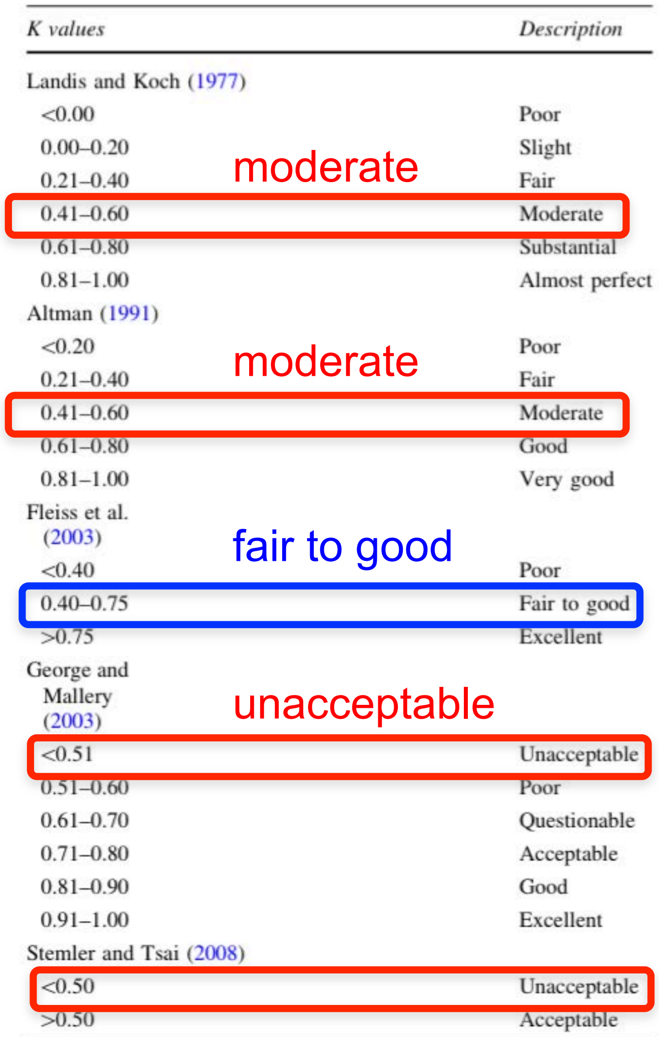


Table 13
Kappa statistic for the amount of agreement between *F* and *P* scores.

	Total sample
	(1)
F and *P*, linear weight kappa	0.54 (18.11)**
F and *P*, VQR weighted kappa	0.54 (17.29)**



[29] Landis and Koch (1977) characterize the range of values 0–0.20 as “slight agreement”, 0.21–0.40 as “fair agreement”, 0.41–0.60 as “moderate agreement”, 0.61–0.80 as “substantial agreement”, and 0.81–1 as “almost perfect agreement”. These guidelines are somewhat arbitrary and by no means universally accepted. Fleiss (1981) for instance characterizes kappas over 0.75 as “excellent”, 0.40 to 0.75 as “fair to good”, and below 0.40 as “poor”. Kappa has also been shown to increase with the number of classes (only 4 in our case). Since the most common scales to subjectively assess the value of kappa mention “adequate” and “fair to good”, these are the terms that we use in the paper to convey the meaning of the statistic when commenting the estimated kappas.



K values	*Description*
Landis and Koch (1977)	
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect
Altman (1991)	
<0.20	Poor
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.00	Very good
Fleiss et al. (2003)	
<0.40	Poor
0.40–0.75	Fair to good
>0.75	Excellent
George and Mallery (2003)	
<0.51	Unacceptable
0.51–0.60	Poor
0.61–0.70	Questionable
0.71–0.80	Acceptable
0.81–0.90	Good
0.91–1.00	Excellent
Stemler and Tsai (2008)	
<0.50	Unacceptable
>0.50	Acceptable

moderate

moderate

fair to good

unacceptable

unacceptable

E negli altri GEV come va?

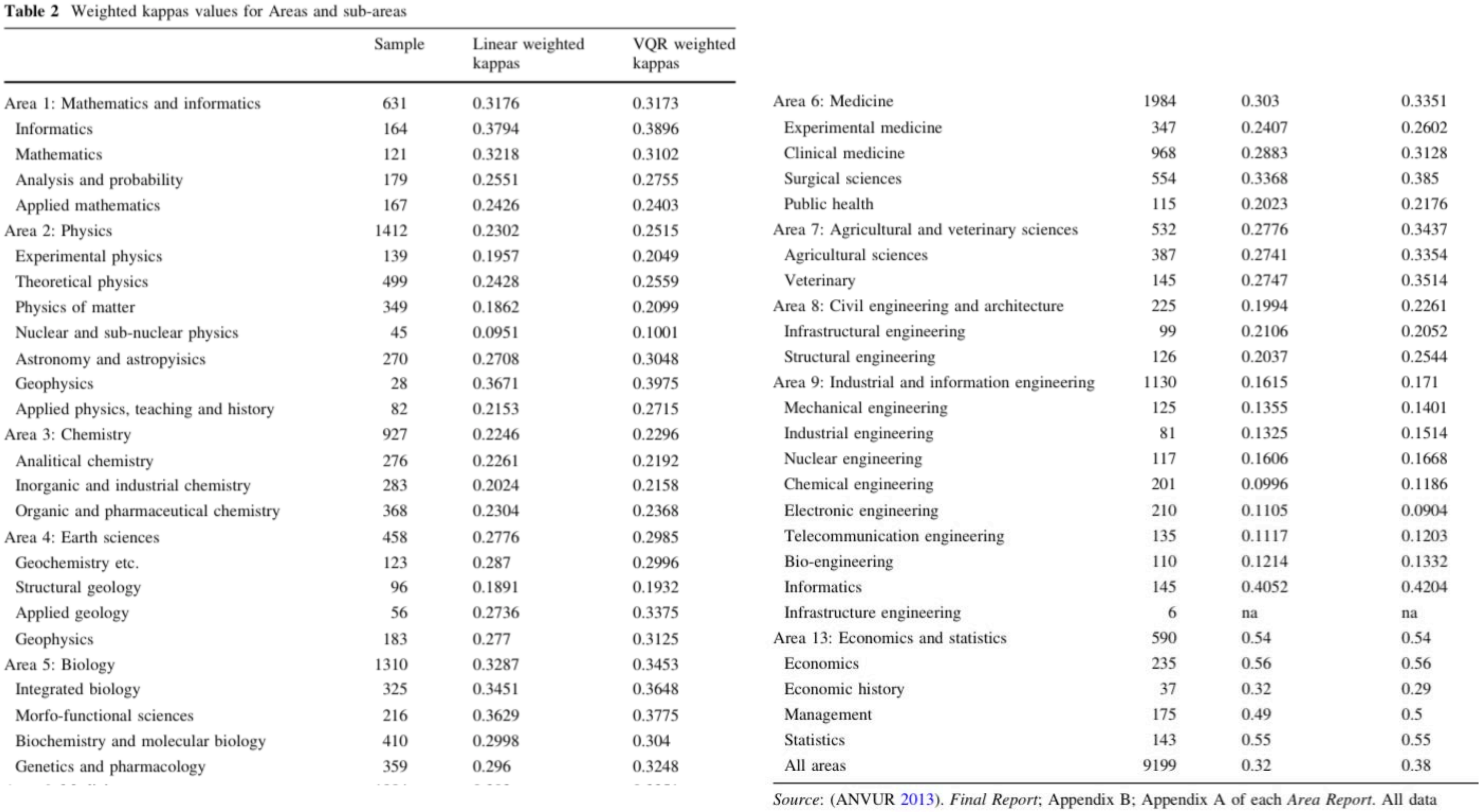
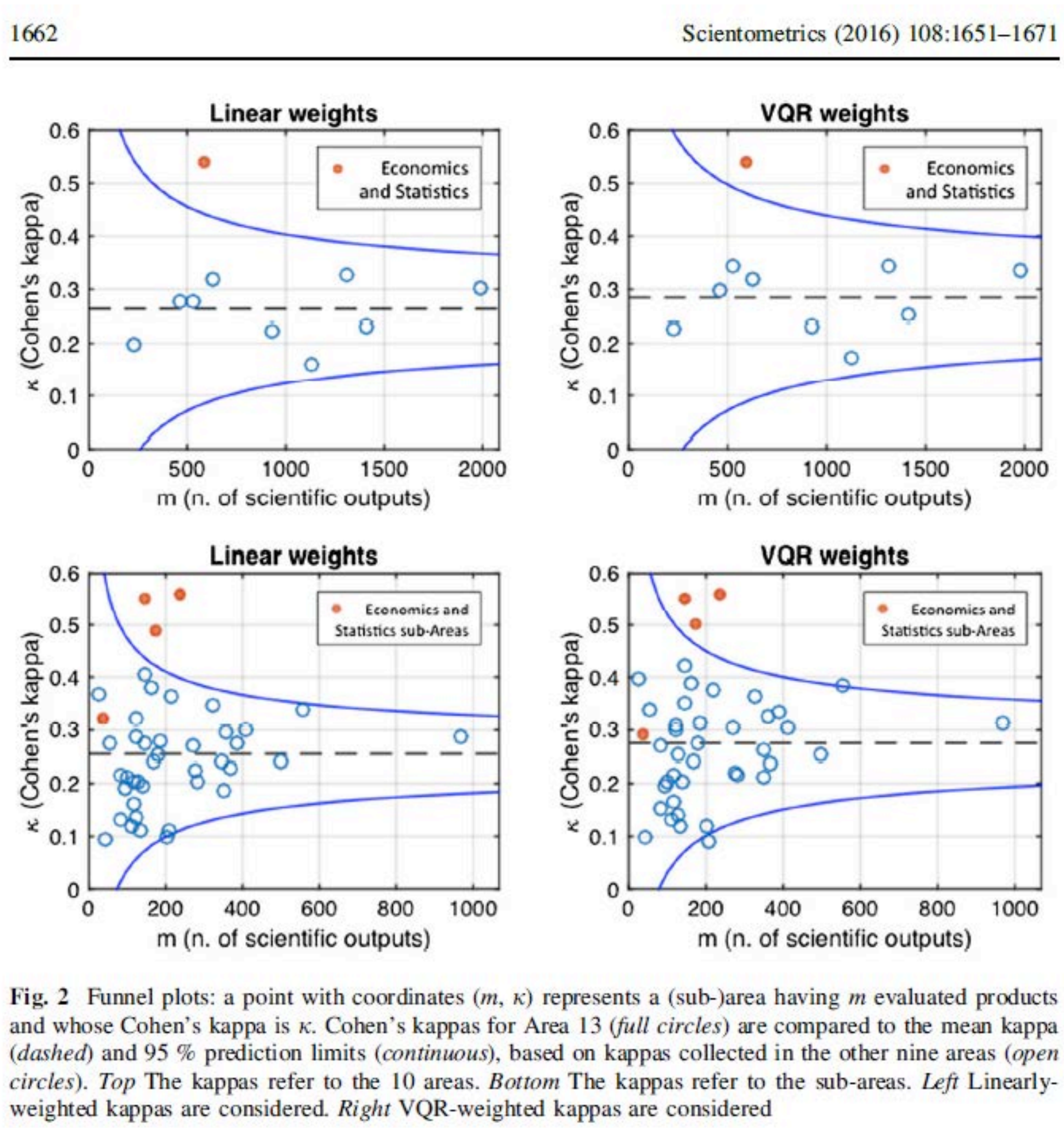


Table 2 Weighted kappas values for Areas and sub-areas

	Sample	Linear weighted kappas	VQR weighted kappas
Area 1: Mathematics and informatics	631	0.3176	0.3173
Informatics	164	0.3794	0.3896
Mathematics	121	0.3218	0.3102
Analysis and probability	179	0.2551	0.2755
Applied mathematics	167	0.2426	0.2403
Area 2: Physics	1412	0.2302	0.2515
Experimental physics	139	0.1957	0.2049
Theoretical physics	499	0.2428	0.2559
Physics of matter	349	0.1862	0.2099
Nuclear and sub-nuclear physics	45	0.0951	0.1001
Astronomy and astropyisics	270	0.2708	0.3048
Geophysics	28	0.3671	0.3975
Applied physics, teaching and history	82	0.2153	0.2715
Area 3: Chemistry	927	0.2246	0.2296
Analitical chemistry	276	0.2261	0.2192
Inorganic and industrial chemistry	283	0.2024	0.2158
Organic and pharmaceutical chemistry	368	0.2304	0.2368
Area 4: Earth sciences	458	0.2776	0.2985
Geochemistry etc.	123	0.287	0.2996
Structural geology	96	0.1891	0.1932
Applied geology	56	0.2736	0.3375
Geophysics	183	0.277	0.3125
Area 5: Biology	1310	0.3287	0.3453
Integrated biology	325	0.3451	0.3648
Morfo-functional sciences	216	0.3629	0.3775
Biochemistry and molecular biology	410	0.2998	0.304
Genetics and pharmacology	359	0.296	0.3248
Area 6: Medicine	1984	0.303	0.3351
Experimental medicine	347	0.2407	0.2602
Clinical medicine	968	0.2883	0.3128
Surgical sciences	554	0.3368	0.385
Public health	115	0.2023	0.2176
Area 7: Agricultural and veterinary sciences	532	0.2776	0.3437
Agricultural sciences	387	0.2741	0.3354
Veterinary	145	0.2747	0.3514
Area 8: Civil engineering and architecture	225	0.1994	0.2261
Infrastructural engineering	99	0.2106	0.2052
Structural engineering	126	0.2037	0.2544
Area 9: Industrial and information engineering	1130	0.1615	0.171
Mechanical engineering	125	0.1355	0.1401
Industrial engineering	81	0.1325	0.1514
Nuclear engineering	117	0.1606	0.1668
Chemical engineering	201	0.0996	0.1186
Electronic engineering	210	0.1105	0.0904
Telecommunication engineering	135	0.1117	0.1203
Bio-engineering	110	0.1214	0.1332
Informatics	145	0.4052	0.4204
Infrastructure engineering	6	na	na
Area 13: Economics and statistics	590	0.54	0.54
Economics	235	0.56	0.56
Economic history	37	0.32	0.29
Management	175	0.49	0.5
Statistics	143	0.55	0.55
All areas	9199	0.32	0.38

Source: (ANVUR 2013). *Final Report*; Appendix B; Appendix A of each *Area Report*. All data



1662 Scientometrics (2016) 108:1651–1671

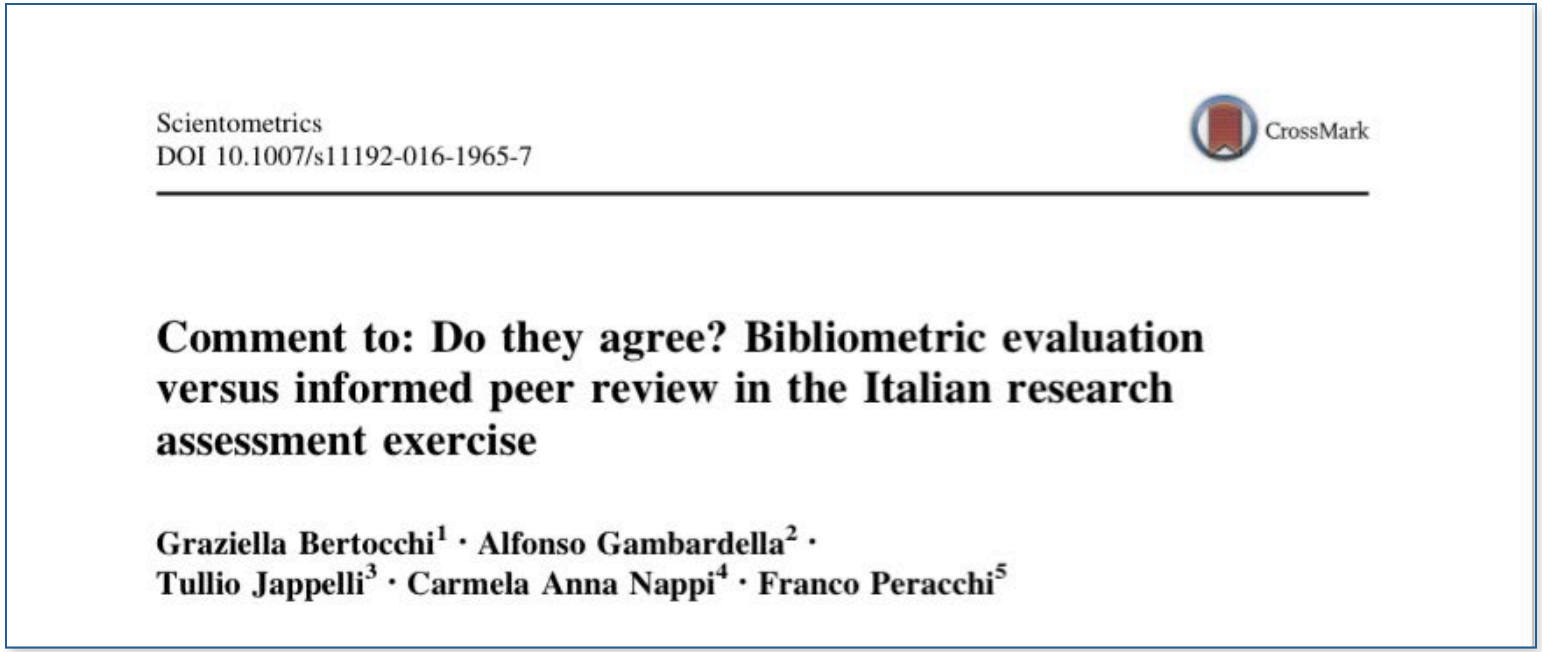
Fig. 2 Funnel plots: a point with coordinates (*m*, *κ*) represents a (sub-)area having *m* evaluated products and whose Cohen's kappa is *κ*. Cohen's kappas for Area 13 (*full circles*) are compared to the mean kappa (*dashed*) and 95 % prediction limits (*continuous*), based on kappas collected in the other nine areas (*open circles*). *Top* The kappas refer to the 10 areas. *Bottom* The kappas refer to the sub-areas. *Left* Linearly-weighted kappas are considered. *Right* VQR-weighted kappas are considered

Cohen's kappa for Economy and Statistics: a statistical anomaly?

Baccini e De Nicolao:
Area 13, “a fatally flawed experiment”

- random sampling took into account authors’ requests to be evaluated by peer review;
- the referees might have known that they were part of the experiment;
- the referees might have known the precise merit class in which each article was classified by using bibliometrics;
- the synthesis of the two referee’s judgments was defined by a Consensus Group composed by (at least) two panel members;
- the panel members forming the Consensus Groups knew that their final judgment would be used for the experiment;
- at least 53 % of the IR evaluations was not expressed by referees, but directly by the Area 13 panelists.

For these reasons, results reached for Area 13 have to be considered as fatally flawed by virtue of the protocol modifications introduced by the area panel



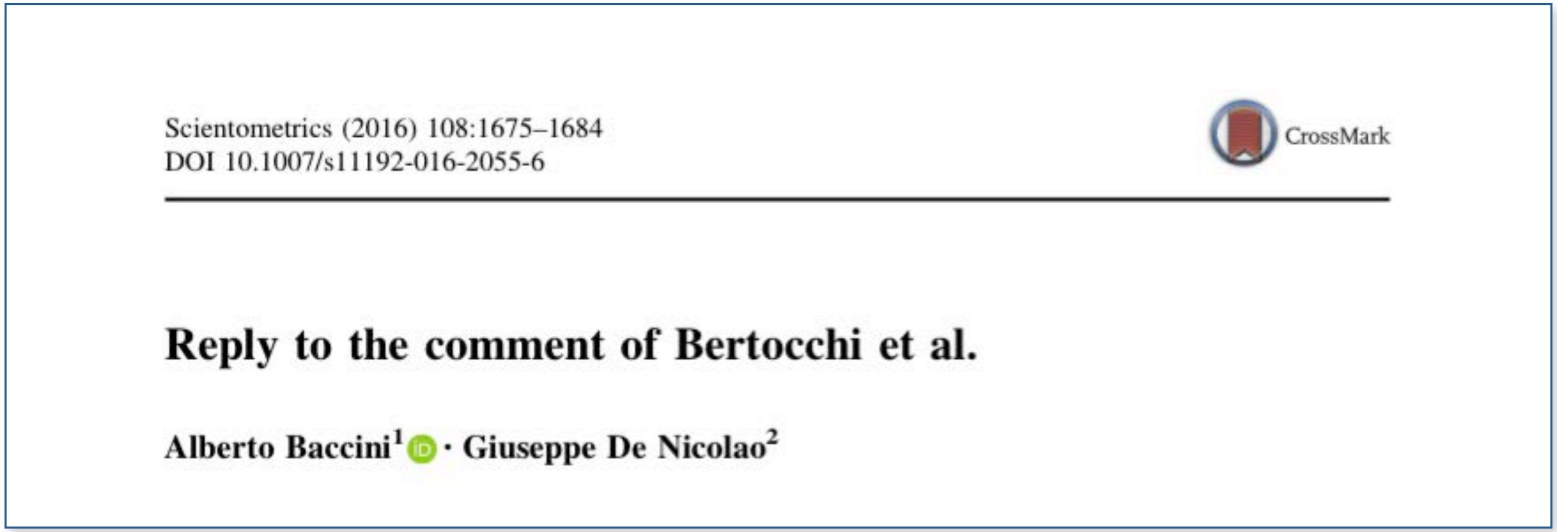
Scientometrics
DOI 10.1007/s11192-016-1965-7

CrossMark

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

Graziella Bertocchi[1] · Alfonso Gambardella[2] · Tullio Jappelli[3] · Carmela Anna Nappi[4] · Franco Peracchi[5]

Many of the points raised by Baccini and De Nicolao (henceforth BD) were already addressed in the RP paper. Other points are either incorrect or not supported by evidence.



Scientometrics (2016) 108:1675–1684
DOI 10.1007/s11192-016-2055-6

CrossMark

Reply to the comment of Bertocchi et al.

Alberto Baccini[1] · Giuseppe De Nicolao[2]

Bertocchi et al.’s comment dismiss our explanation and suggest that the difference was due to ‘‘differences in the evaluation processes between Area 13 and other areas’’. In addition, they state that all our five claims about Area 13 experiment protocol ‘‘are either incorrect or not based on any evidence’’. Based on textual evidence drawn from ANVUR official reports, we show that: (1) none of the four differences listed by Bertocchi et al. is peculiar of Area 13; (2) their five arguments contesting our claims about the experiment protocol are all contradicted by official records of the experiment itself.

5. Concordanza
o fallacia statistica?

Research Evaluation (2015) pp. 1–14 doi:10.1093/reseval/rvv008

Evaluating scientific research in Italy: The 2004–10 research evaluation exercise

Alessio Ancaiani[1], Alberto F. Anfossi[1,2], Anna Barbara[1,3], Sergio Benedetto[1], Brigida Blasi[1], Valentina Carletti[1], Tindaro Cicero[1], Alberto Ciolfi[1], Filippo Costa[1,4], Giovanna Colizza[1], Marco Costantini[1,3], Fabio di Cristina[1], Antonio Ferrara[1], Rosa M. Lacatena[1], Marco Malgarini[1,*], Irene Mazzotta[1], Carmela A. Nappi[1], Sandra Romagnosi[1] and Serena Sileoni[1]

[1] Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR), Via Ippolito Nievo 35 - 00153 Rome, Italy, [2] Compagnia di San Paolo Sistema Torino, Piazza Bernini 5, IT-10138 Turin, Italy, [3] Gabriele D'Annunzio Chieti-Pescara University Via dei Vestini, 31 - 66013 Chieti Scalo, Italy and [4] Department of Information Engineering, Pisa University, Via Caruso 16 - 56122 Pisa, Italy

**Corresponding author. Email: marco.malgarini@anvur.it*

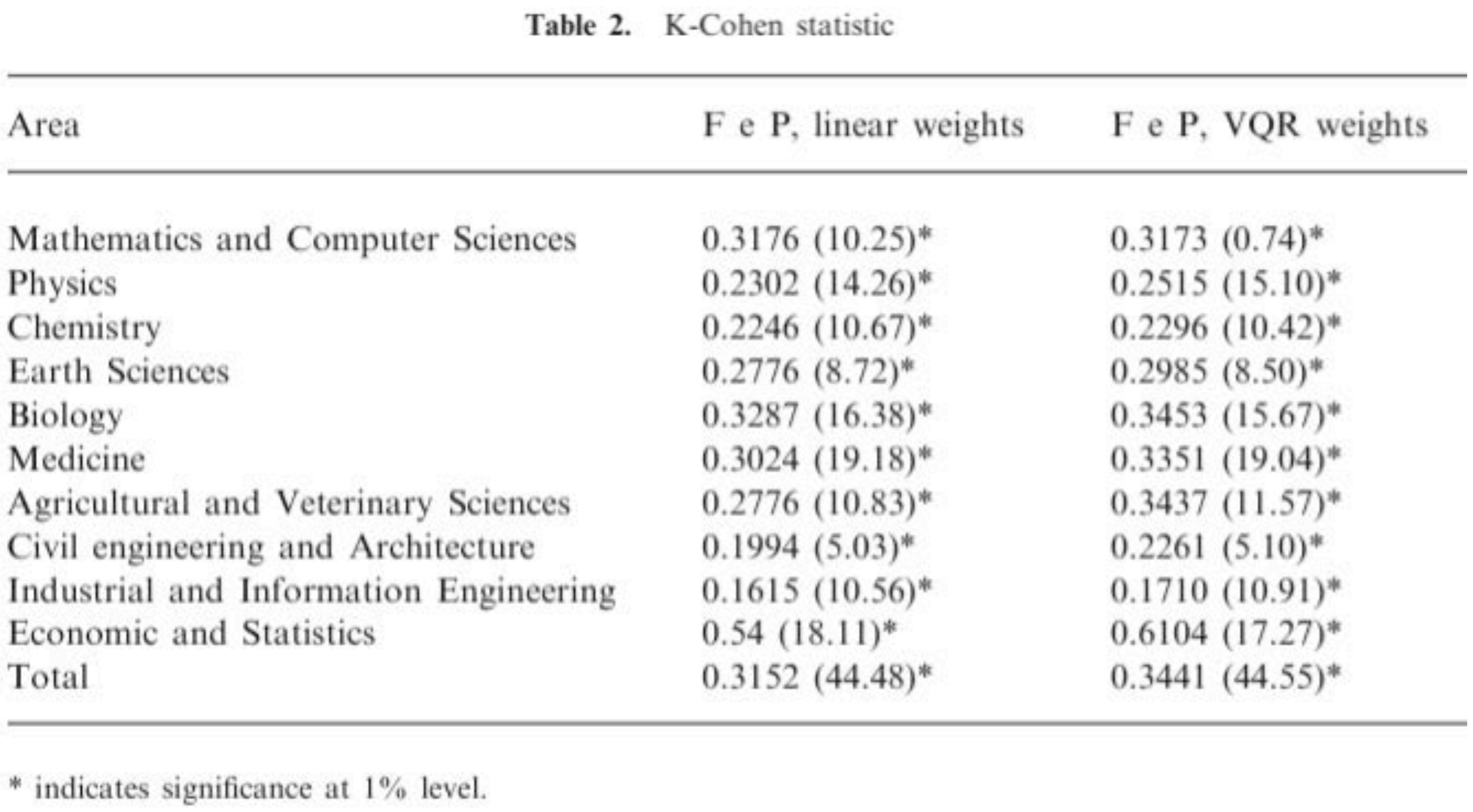


Table 2. K-Cohen statistic

Area	F e P, linear weights	F e P, VQR weights
Mathematics and Computer Sciences	0.3176 (10.25)*	0.3173 (0.74)*
Physics	0.2302 (14.26)*	0.2515 (15.10)*
Chemistry	0.2246 (10.67)*	0.2296 (10.42)*
Earth Sciences	0.2776 (8.72)*	0.2985 (8.50)*
Biology	0.3287 (16.38)*	0.3453 (15.67)*
Medicine	0.3024 (19.18)*	0.3351 (19.04)*
Agricultural and Veterinary Sciences	0.2776 (10.83)*	0.3437 (11.57)*
Civil engineering and Architecture	0.1994 (5.03)*	0.2261 (5.10)*
Industrial and Information Engineering	0.1615 (10.56)*	0.1710 (10.91)*
Economic and Statistics	0.54 (18.11)*	0.6104 (17.27)*
Total	0.3152 (44.48)*	0.3441 (44.55)*

* indicates significance at 1% level.

«K is always ***statistically different from zero****, showing that there is a* ***fundamental agreement*** *among the two distributions which* ***may not be attributed to mere chance****, regardless of the weight used to calculate the differences among the two distributions. The value of K ranges from 0.16 to 0.61 depending on the area and weights, being on average equal to 0.32, a value that is usually considered as* ***'poor to fair'*** *in the literature (Landis and Koch 1977).»*

Therefore:

“results of the analysis relative to the degree of concordance and systematic difference may be considered to ***validate the general approach of combining peer review and bibliometric methods****” (Ancaiani et al. 2015)*

Is this true?

Una nozione insegnata in tutti i corsi di statistica di base: la differenza tra ***statistical*** e ***practical*** significance

Researchers want to test a new medication that claims to raise IQs to genius levels (175+). In the population, the average IQ is 100. A sample of 40 individuals has a mean IQ of 110 with a standard deviation of 15.

A t-test is performed and the null hypothesis is rejected. It is concluded that the medication raises IQ.

But the medication claimed to raise IQs to genius (175+) levels. Even though we found statistical significance, the medication does not meet the practical value it claimed to. It lacks practical significance.

Statistical vs Practical Significance

Practical Significance

- Statistically significant results (rejecting the null hypothesis) are not always of practical significance
  - This is more likely to happen when the sample size gets very large
- Practically significant results might be found to be statistically insignificant (failing to reject the null hypothesis)
  - This is more likely to happen when the sample size is relatively small

Practical Significance

- Practical (or clinical) significance asks the larger question about differences
  - "Are the differences between samples big enough to have real meaning?"
- Although men and women undoubtedly have different IQs, is that difference large enough to have some practical implication

Statistical Versus Practical Significance

For a reasonably large sample size n, this μ would lead to an x̄ value near 101, so we would not want this sample evidence to argue strongly for rejection of H_0 when x̄ = 101 is observed.

For various sample sizes, Table 8.1 records both the P-value when x̄ = 101 and also the probability of not rejecting H_0 at level .01 when μ = 101.

Products produced by the machine (following the repair)

Sample (400 products)

64 defective

Defective

p

H_0: p = .20

H_a: p < .20

Conditions are met

p̂ = 64/400 = .16

z = -2

p-value = .023

Conclusion: H_0 can be rejected.

Statistical vs practical significance

Example

- large keyboard typing: 30.2 chars/minute
- small keyboard typing: 29.9
- differences statistically significant

But

- people generally type short strings
- time savings not critical
- screen space more important than time savings

Recommendation

- use small keyboard

Statistical vs Clinical or Practical Significance

Will G Hopkins

Auckland University of Technology

Auckland, NZ

- Statistical significance
  - P values and null hypotheses
- Confidence limits
  - Precision of estimation
- Clinical or practical significance
  - Probabilities of benefit and harm
  - Examples

Statistical Significance

Statistical vs. Practical Significance

Statistical vs. Practical Significance ?

Experimental significance

Statistical significance

Practical significance

"Are the effects observed due to the changing levels of the factors real or due to random chance?"

"If a factor does have a statistically significant effect, then is the degree of effect practically useful given what we are trying to accomplish."

Statistical vs Practical Significance

- Example: Suppose a weight loss program recruits 10,000 people for a randomized experiment.
- A difference in average weight loss of only 0.5 lbs could be found to be statistically significant
- Suppose the experiment lasted for a year. Is a loss of ½ a pound practically significant?

Statistical versus Practical Significance

- When we reject a null hypothesis it is usual to say the result is statistically significant at the chosen level of significance
- But should also always consider the practical significance of the magnitude of the difference between the estimate (of the population characteristic) and what the null hypothesis states that to be

Statistical vs Practical Significance

- With small sample sizes, even large differences or effects may not be significant
- With large sample sizes, even a very small difference or effect can be significant
- A statistically significant result is not always practically significant, especially with large sample sizes

Practical Significance vs. Statistical Significance

- An increase in the average SAT-M score from 475 to 478 is of little importance in seeking admission to college, but a large enough sample size will always declare very small effects statistically significant.
- A confidence interval provides information about the size of the effect and should always be reported. The two-sided 95% confidence intervals for the SAT coaching problem are

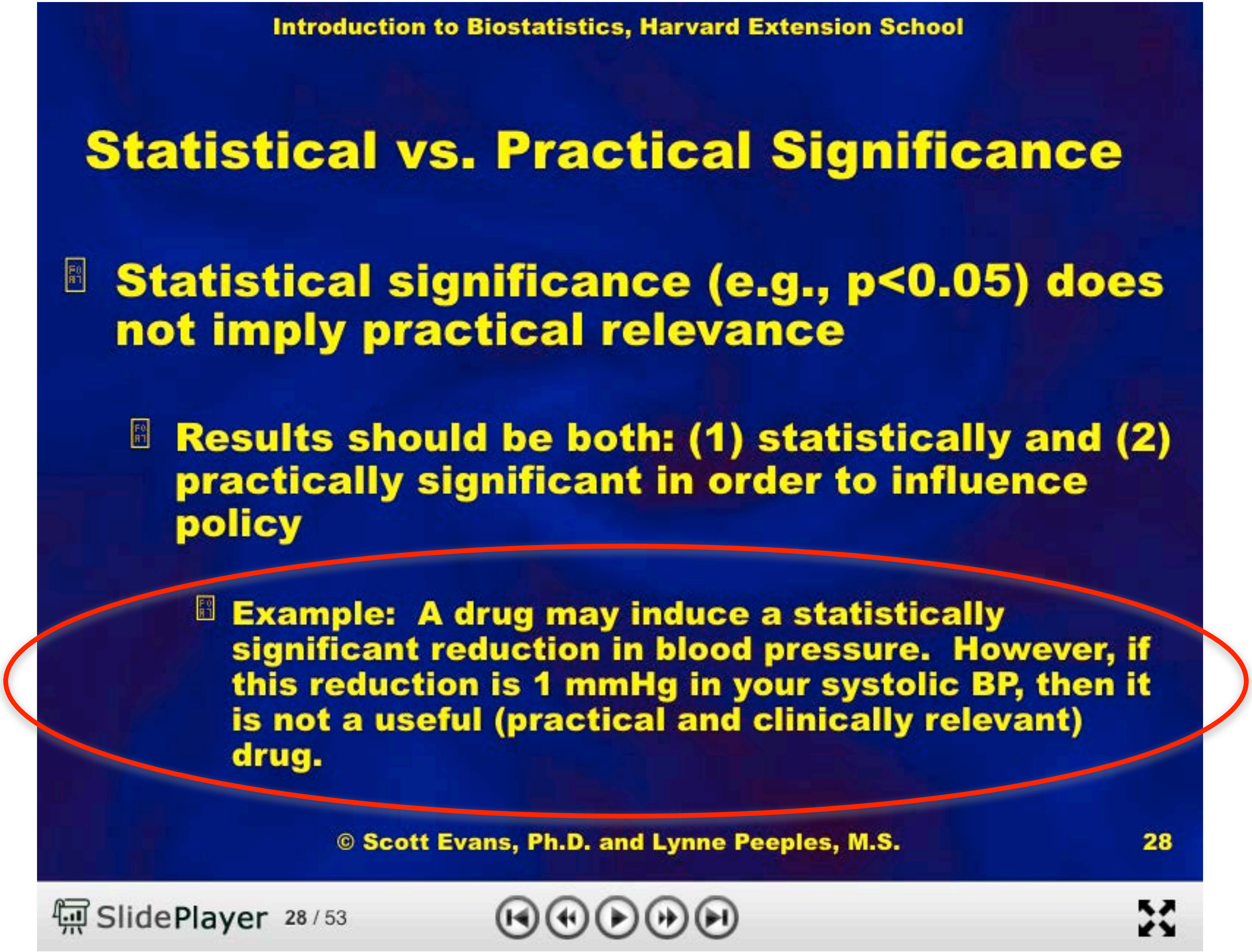
A COMMON MISTAKES

Not everything can be quantified!

Statistical vs Practical Significance

Example: Suppose a weight loss program recruits 10,000 people for a randomized experiment.

- A difference in average weight loss of only 0.5 lbs could be found to be statistically significant
- Suppose the experiment lasted for a year. Is a loss of ½ a pound practically significant?



Introduction to Biostatistics, Harvard Extension School

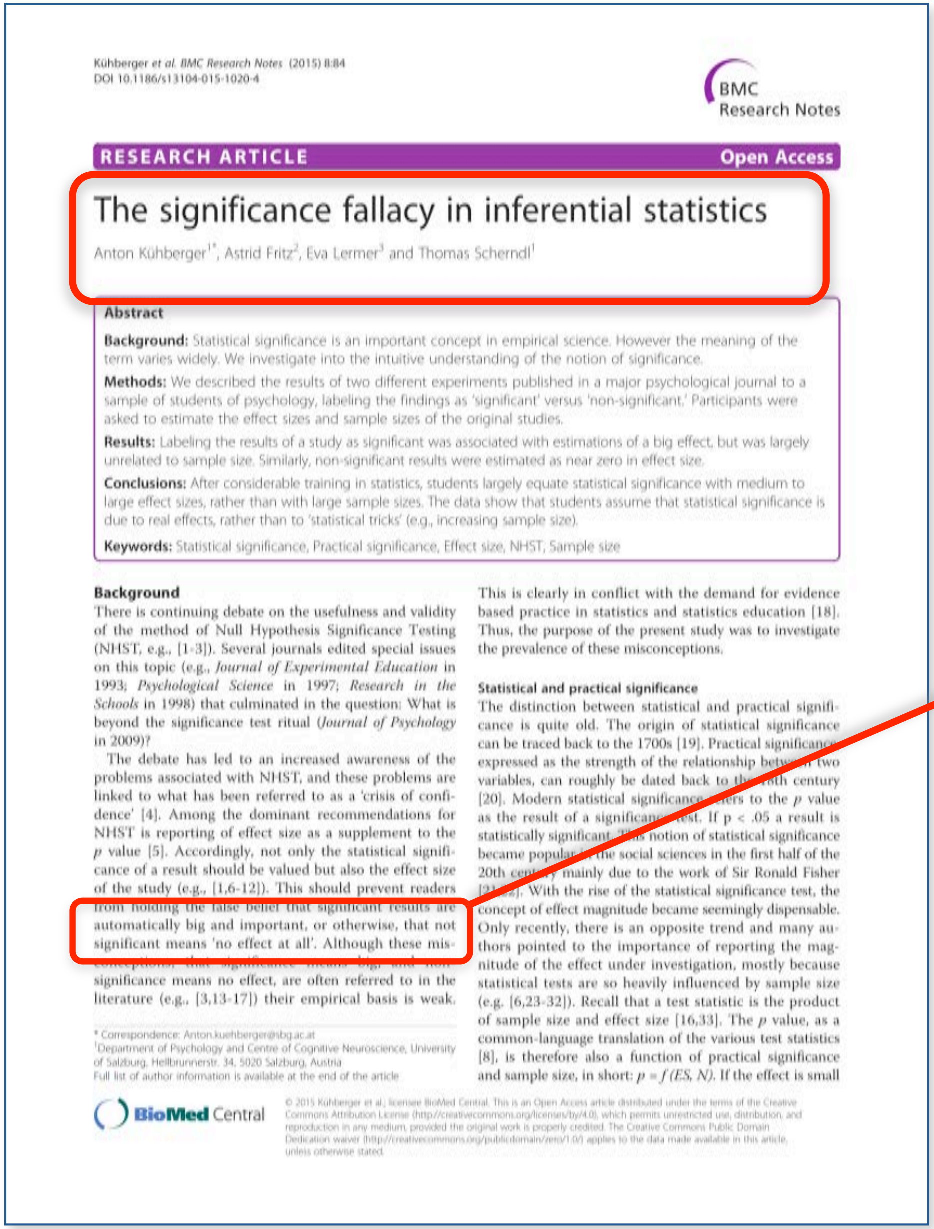
Statistical vs. Practical Significance

- Statistical significance (e.g., p<0.05) does not imply practical relevance
  - Results should be both: (1) statistically and (2) practically significant in order to influence policy
    - Example: A drug may induce a statistically significant reduction in blood pressure. However, if this reduction is 1 mmHg in your systolic BP, then it is not a useful (practical and clinically relevant) drug.

© Scott Evans, Ph.D. and Lynne Peeples, M.S.

28

The significance fallacy



Kühberger *et al. BMC Research Notes* (2015) 8:84
DOI 10.1186/s13104-015-1020-4

BMC
Research Notes

RESEARCH ARTICLE Open Access

The significance fallacy in inferential statistics

Anton Kühberger[1*], Astrid Fritz[2], Eva Lermer[3] and Thomas Scherndl[1]

Abstract

Background: Statistical significance is an important concept in empirical science. However the meaning of the term varies widely. We investigate into the intuitive understanding of the notion of significance.

Methods: We described the results of two different experiments published in a major psychological journal to a sample of students of psychology, labeling the findings as 'significant' versus 'non-significant.' Participants were asked to estimate the effect sizes and sample sizes of the original studies.

Results: Labeling the results of a study as significant was associated with estimations of a big effect, but was largely unrelated to sample size. Similarly, non-significant results were estimated as near zero in effect size.

Conclusions: After considerable training in statistics, students largely equate statistical significance with medium to large effect sizes, rather than with large sample sizes. The data show that students assume that statistical significance is due to real effects, rather than to 'statistical tricks' (e.g., increasing sample size).

Keywords: Statistical significance, Practical significance, Effect size, NHST, Sample size

Background

There is continuing debate on the usefulness and validity of the method of Null Hypothesis Significance Testing (NHST, e.g., [1-3]). Several journals edited special issues on this topic (e.g., *Journal of Experimental Education* in 1993; *Psychological Science* in 1997; *Research in the Schools* in 1998) that culminated in the question: What is beyond the significance test ritual (*Journal of Psychology* in 2009)?

The debate has led to an increased awareness of the problems associated with NHST, and these problems are linked to what has been referred to as a 'crisis of confidence' [4]. Among the dominant recommendations for NHST is reporting of effect size as a supplement to the *p* value [5]. Accordingly, not only the statistical significance of a result should be valued but also the effect size of the study (e.g., [1,6-12]). This should prevent readers from holding the false belief that significant results are automatically big and important, or otherwise, that not significant means 'no effect at all'. Although these misconceptions, that significance means big, and non-significance means no effect, are often referred to in the literature (e.g., [3,13-17]) their empirical basis is weak. This is clearly in conflict with the demand for evidence based practice in statistics and statistics education [18]. Thus, the purpose of the present study was to investigate the prevalence of these misconceptions.

Statistical and practical significance

The distinction between statistical and practical significance is quite old. The origin of statistical significance can be traced back to the 1700s [19]. Practical significance, expressed as the strength of the relationship between two variables, can roughly be dated back to the 19th century [20]. Modern statistical significance refers to the *p* value as the result of a significance test. If $p < .05$ a result is statistically significant. This notion of statistical significance became popular in the social sciences in the first half of the 20th century mainly due to the work of Sir Ronald Fisher [21,22]. With the rise of the statistical significance test, the concept of effect magnitude became seemingly dispensable. Only recently, there is an opposite trend and many authors pointed to the importance of reporting the magnitude of the effect under investigation, mostly because statistical tests are so heavily influenced by sample size (e.g. [6,23-32]). Recall that a test statistic is the product of sample size and effect size [16,33]. The *p* value, as a common-language translation of the various test statistics [8], is therefore also a function of practical significance and sample size, in short: $p = f\,(ES, N)$. If the effect is small

[*] Correspondence: Anton.kuehberger@sbg.ac.at
[1] Department of Psychology and Centre of Cognitive Neuroscience, University of Salzburg, Hellbrunnerstr. 34, 5020 Salzburg, Austria
Full list of author information is available at the end of the article

BioMed Central

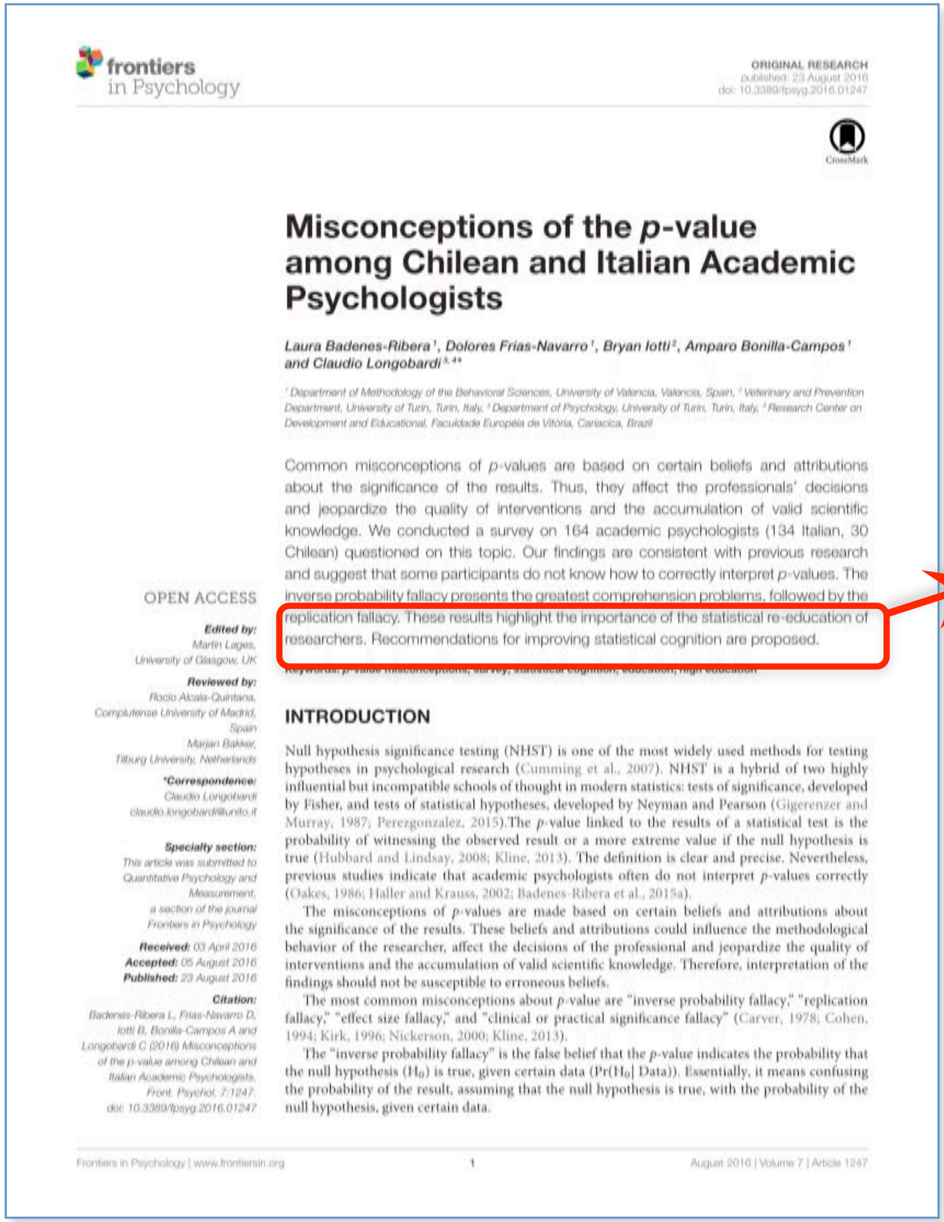
the false belief that [statistically] significant results are automatically big and important

Una citazione riferita proprio alla kappa di Cohen

Statistical significance “is generally of little practical value, since a relatively low value of kappa can yield a significant result. In other words, a value such as k = 0.41 (in spite of the fact that is statistically significant) may be deemed by a researcher to be too low a level of reliability (i.e. degree of agreement) to be utilized within a practical context” (Sheskin 2003).

“the results reported by Ancaiani et al ***. do not support a good concordance between peer review and bibliometrics****. [...] On the basis of these data, the conclusion that it is possible to use both technique as interchangeable in a research assessment exercise appears to be* ***unsound****.” (Baccini and De Nicolao 2017)*

Statistical re-education needed



frontiers in Psychology

ORIGINAL RESEARCH
published: 23 August 2016
doi: 10.3389/fpsyg.2016.01247

Misconceptions of the *p*-value among Chilean and Italian Academic Psychologists

Laura Badenes-Ribera[1], *Dolores Frias-Navarro*[1], *Bryan Iotti*[2], *Amparo Bonilla-Campos*[1] *and Claudio Longobardi*[3,4]*

[1] *Department of Methodology of the Behavioral Sciences, University of Valencia, Valencia, Spain,* [2] *Veterinary and Prevention Department, University of Turin, Turin, Italy,* [3] *Department of Psychology, University of Turin, Turin, Italy,* [4] *Research Center on Development and Educational, Faculdade Européia de Vitória, Cariacica, Brazil*

Common misconceptions of *p*-values are based on certain beliefs and attributions about the significance of the results. Thus, they affect the professionals' decisions and jeopardize the quality of interventions and the accumulation of valid scientific knowledge. We conducted a survey on 164 academic psychologists (134 Italian, 30 Chilean) questioned on this topic. Our findings are consistent with previous research and suggest that some participants do not know how to correctly interpret *p*-values. The inverse probability fallacy presents the greatest comprehension problems, followed by the replication fallacy. These results highlight the importance of the statistical re-education of researchers. Recommendations for improving statistical cognition are proposed.

Keywords: *p*-value misconceptions, survey, statistical cognition, education, high education

OPEN ACCESS

Edited by:
Martin Lages,
University of Glasgow, UK

Reviewed by:
Rocio Alcala-Quintana,
Complutense University of Madrid, Spain
Marjan Bakker,
Tilburg University, Netherlands

****Correspondence:***
Claudio Longobardi
claudio.longobardi@unito.it

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal Frontiers in Psychology

Received: *03 April 2016*
Accepted: *05 August 2016*
Published: *23 August 2016*

Citation:
Badenes-Ribera L, Frias-Navarro D, Iotti B, Bonilla-Campos A and Longobardi C (2016) Misconceptions of the p-value among Chilean and Italian Academic Psychologists. Front. Psychol. 7:1247. doi: 10.3389/fpsyg.2016.01247

INTRODUCTION

Null hypothesis significance testing (NHST) is one of the most widely used methods for testing hypotheses in psychological research (Cumming et al., 2007). NHST is a hybrid of two highly influential but incompatible schools of thought in modern statistics: tests of significance, developed by Fisher, and tests of statistical hypotheses, developed by Neyman and Pearson (Gigerenzer and Murray, 1987; Perezgonzalez, 2015).The *p*-value linked to the results of a statistical test is the probability of witnessing the observed result or a more extreme value if the null hypothesis is true (Hubbard and Lindsay, 2008; Kline, 2013). The definition is clear and precise. Nevertheless, previous studies indicate that academic psychologists often do not interpret *p*-values correctly (Oakes, 1986; Haller and Krauss, 2002; Badenes-Ribera et al., 2015a).

The misconceptions of *p*-values are made based on certain beliefs and attributions about the significance of the results. These beliefs and attributions could influence the methodological behavior of the researcher, affect the decisions of the professional and jeopardize the quality of interventions and the accumulation of valid scientific knowledge. Therefore, interpretation of the findings should not be susceptible to erroneous beliefs.

The most common misconceptions about *p*-value are "inverse probability fallacy," "replication fallacy," "effect size fallacy," and "clinical or practical significance fallacy" (Carver, 1978; Cohen, 1994; Kirk, 1996; Nickerson, 2000; Kline, 2013).

The "inverse probability fallacy" is the false belief that the *p*-value indicates the probability that the null hypothesis (H_0) is true, given certain data ($Pr(H_0| Data)$). Essentially, it means confusing the probability of the result, assuming that the null hypothesis is true, with the probability of the null hypothesis, given certain data.

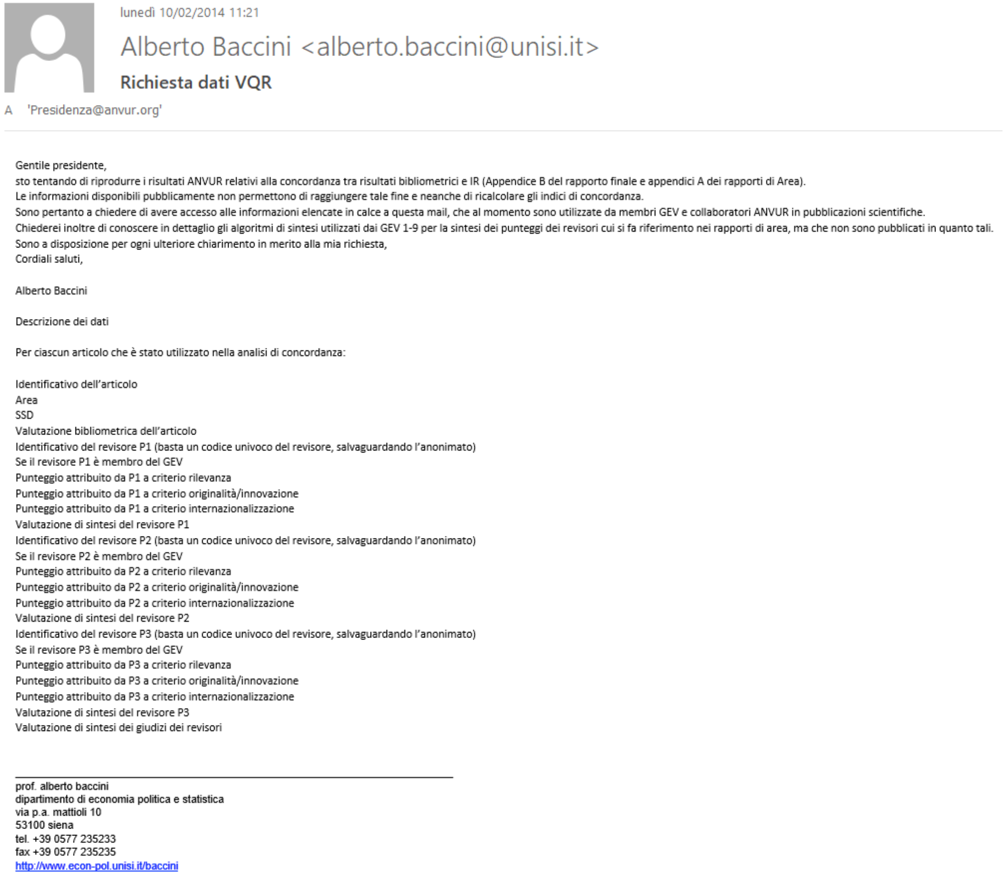
Frontiers in Psychology | www.frontiersin.org 1 August 2016 | Volume 7 | Article 1247

These results highlight the importance of the statistical re-education of researchers

6. Dati chiusi, concordanza non replicabile

Dal 2014 abbiamo tentato di replicare l'esperimento

- ANVUR non fornisce i dati necessari (mail 10/2/2014 a Presidente Fantoni)



lunedì 10/02/2014 11:21

Alberto Baccini <alberto.baccini@unisi.it>

Richiesta dati VQR

A 'Presidenza@anvur.org'

Gentile presidente,
sto tentando di riprodurre i risultati ANVUR relativi alla concordanza tra risultati bibliometrici e IR (Appendice B del rapporto finale e appendici A dei rapporti di Area).
Le informazioni disponibili pubblicamente non permettono di raggiungere tale fine e neanche di ricalcolare gli indici di concordanza.
Sono pertanto a chiedere di avere accesso alle informazioni elencate in calce a questa mail, che al momento sono utilizzate da membri GEV e collaboratori ANVUR in pubblicazioni scientifiche.
Chiederei inoltre di conoscere in dettaglio gli algoritmi di sintesi utilizzati dai GEV 1-9 per la sintesi dei punteggi dei revisori cui si fa riferimento nei rapporti di area, ma che non sono pubblicati in quanto tali.
Sono a disposizione per ogni ulteriore chiarimento in merito alla mia richiesta,
Cordiali saluti,

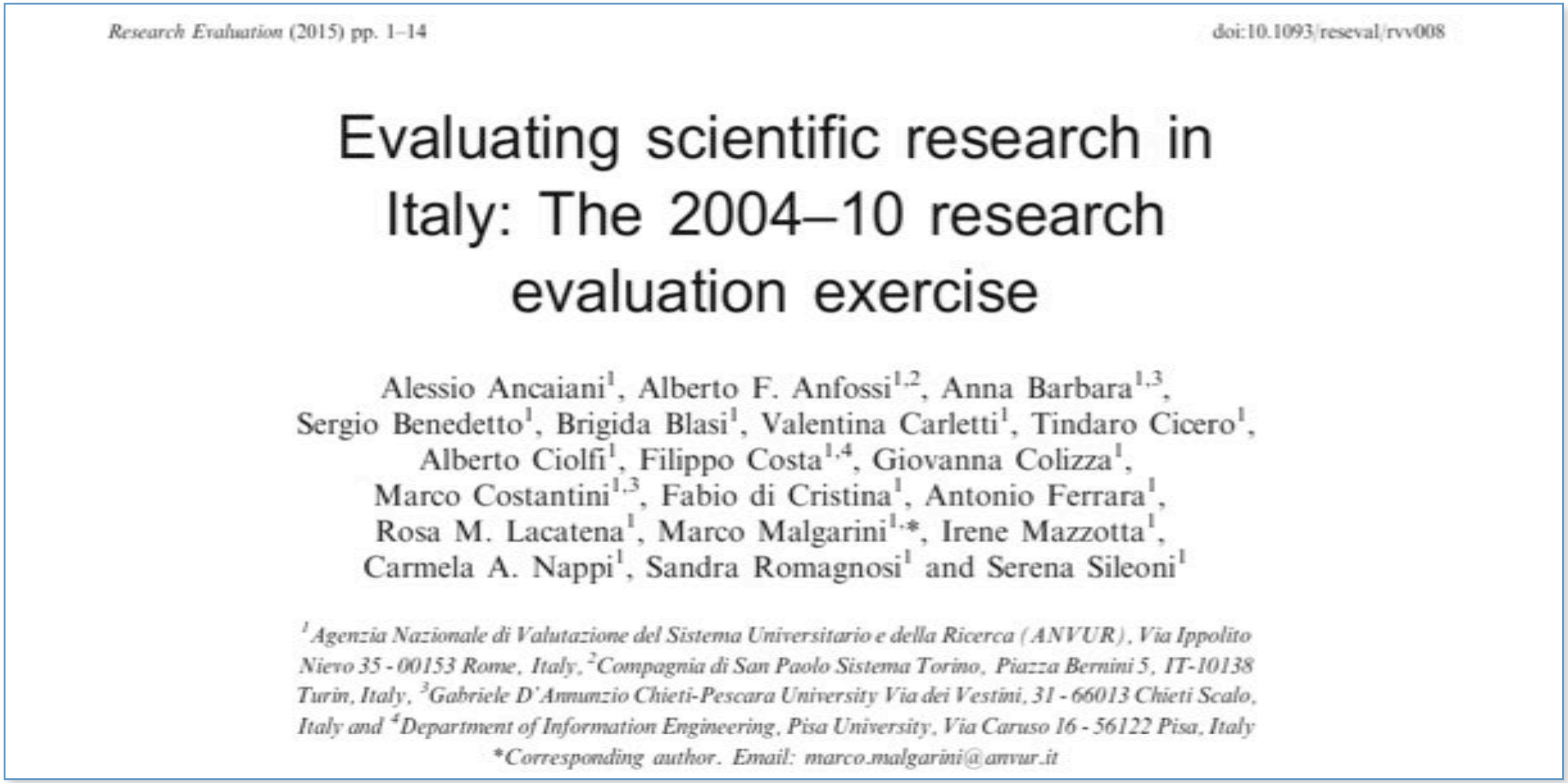
Alberto Baccini

Descrizione dei dati

Per ciascun articolo che è stato utilizzato nella analisi di concordanza:

Identificativo dell'articolo
Area
SSD
Valutazione bibliometrica dell'articolo
Identificativo del revisore P1 (basta un codice univoco del revisore, salvaguardando l'anonimato)
Se il revisore P1 è membro del GEV
Punteggio attribuito da P1 a criterio rilevanza
Punteggio attribuito da P1 a criterio originalità/innovazione
Punteggio attribuito da P1 a criterio internazionalizzazione
Valutazione di sintesi del revisore P1
Identificativo del revisore P2 (basta un codice univoco del revisore, salvaguardando l'anonimato)
Se il revisore P2 è membro del GEV
Punteggio attribuito da P2 a criterio rilevanza
Punteggio attribuito da P2 a criterio originalità/innovazione
Punteggio attribuito da P2 a criterio internazionalizzazione
Valutazione di sintesi del revisore P2
Identificativo del revisore P3 (basta un codice univoco del revisore, salvaguardando l'anonimato)
Se il revisore P3 è membro del GEV
Punteggio attribuito da P3 a criterio rilevanza
Punteggio attribuito da P3 a criterio originalità/innovazione
Punteggio attribuito da P3 a criterio internazionalizzazione
Valutazione di sintesi del revisore P3
Valutazione di sintesi dei giudizi dei revisori

prof. alberto baccini
dipartimento di economia politica e statistica
via p.a. mattioli 10
53100 siena
tel. +39 0577 235233
fax +39 0577 235235
http://www.econ-pol.unisi.it/baccini



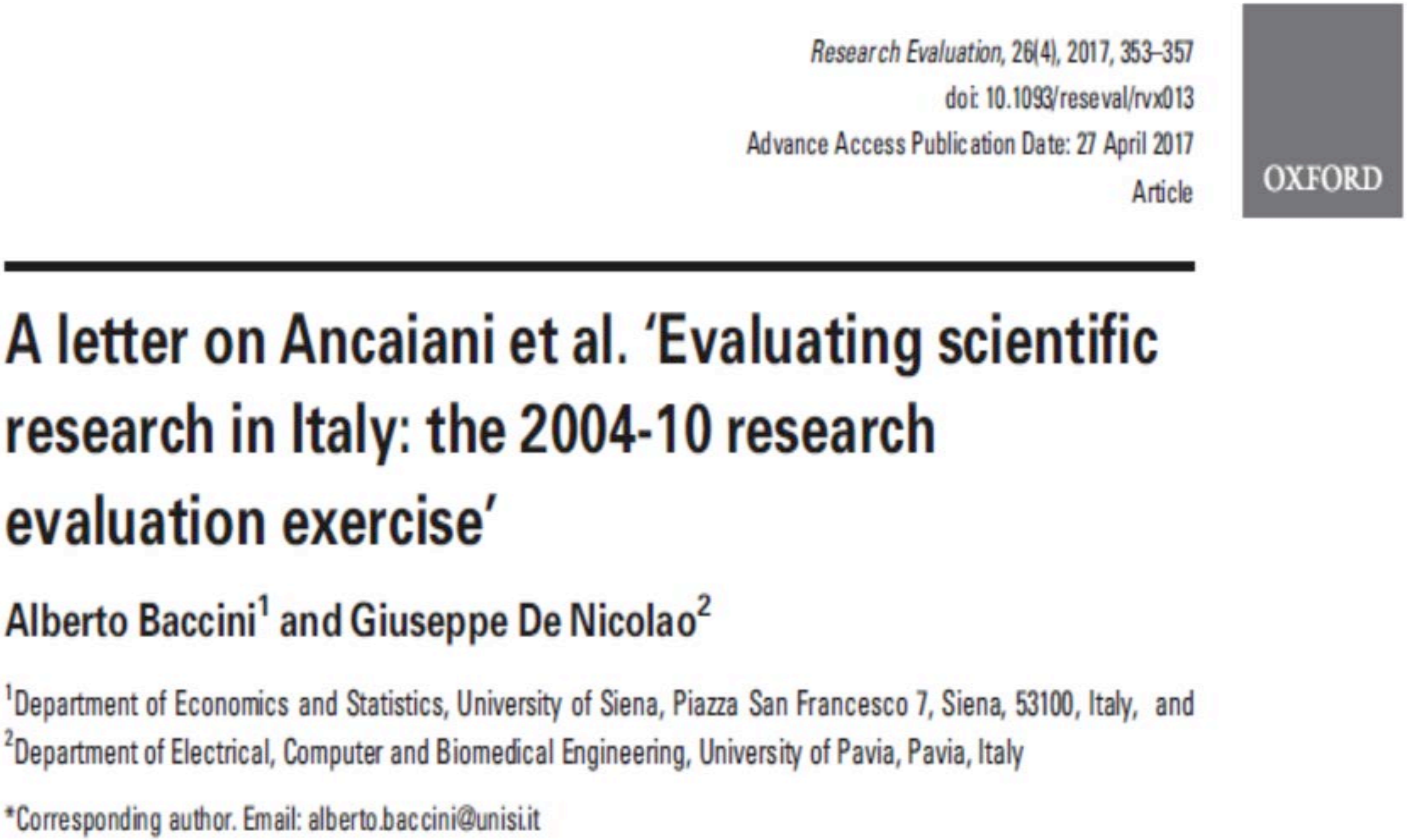
Research Evaluation (2015) pp. 1–14 doi:10.1093/reseval/rvv008

Evaluating scientific research in Italy: The 2004–10 research evaluation exercise

Alessio Ancaiani[1], Alberto F. Anfossi[1,2], Anna Barbara[1,3],
Sergio Benedetto[1], Brigida Blasi[1], Valentina Carletti[1], Tindaro Cicero[1],
Alberto Ciolfi[1], Filippo Costa[1,4], Giovanna Colizza[1],
Marco Costantini[1,3], Fabio di Cristina[1], Antonio Ferrara[1],
Rosa M. Lacatena[1], Marco Malgarini[1,*], Irene Mazzotta[1],
Carmela A. Nappi[1], Sandra Romagnosi[1] and Serena Sileoni[1]

[1] Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR), Via Ippolito Nievo 35 - 00153 Rome, Italy, [2] Compagnia di San Paolo Sistema Torino, Piazza Bernini 5, IT-10138 Turin, Italy, [3] Gabriele D'Annunzio Chieti-Pescara University Via dei Vestini, 31 - 66013 Chieti Scalo, Italy and [4] Department of Information Engineering, Pisa University, Via Caruso 16 - 56122 Pisa, Italy

**Corresponding author. Email: marco.malgarini@anvur.it*



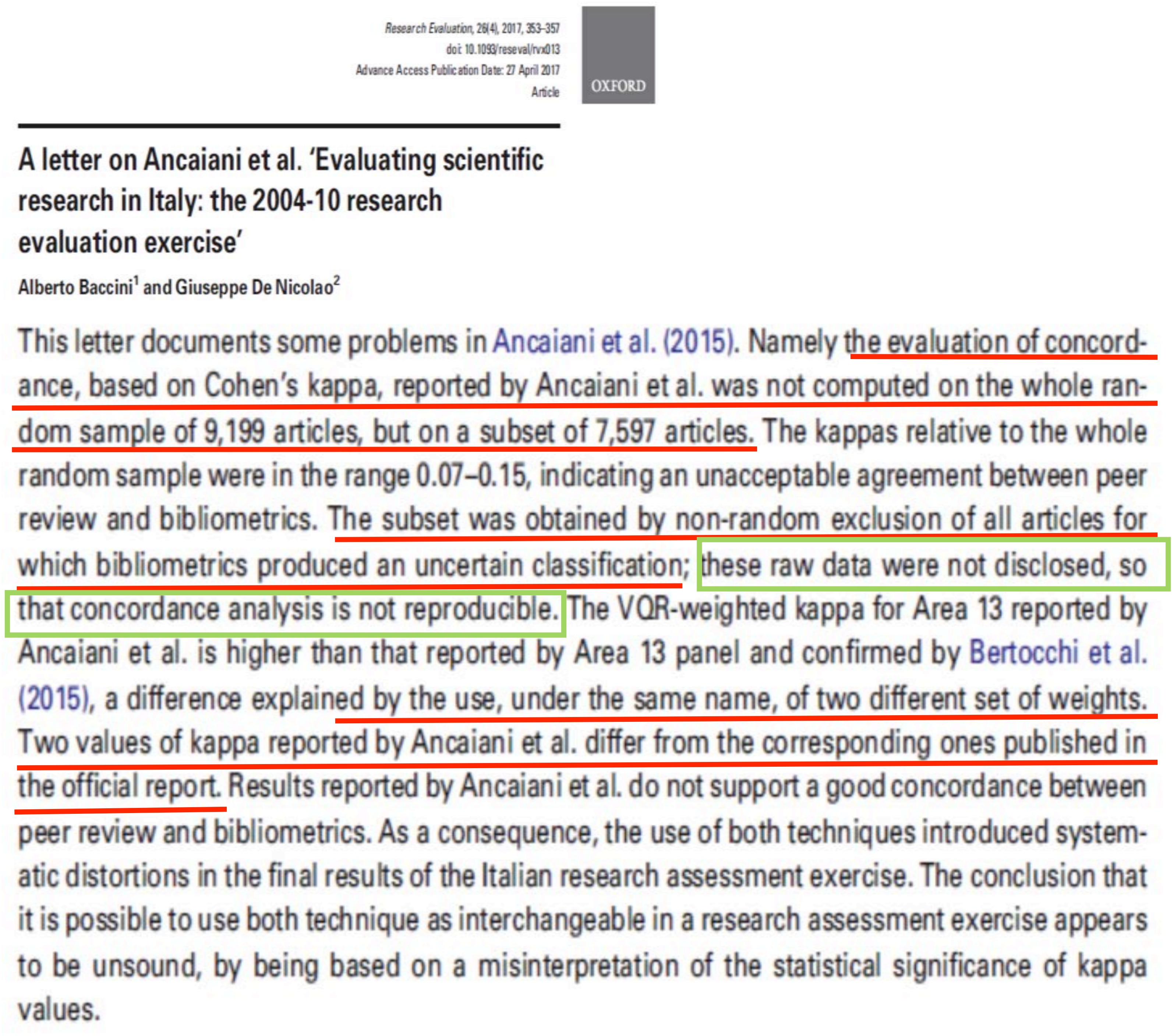
Research Evaluation, 26(4), 2017, 353–357
doi: 10.1093/reseval/rvx013
Advance Access Publication Date: 27 April 2017
Article

A letter on Ancaiani et al. 'Evaluating scientific research in Italy: the 2004-10 research evaluation exercise'

Alberto Baccini[1] and Giuseppe De Nicolao[2]

[1]Department of Economics and Statistics, University of Siena, Piazza San Francesco 7, Siena, 53100, Italy, and
[2]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy

*Corresponding author. Email: alberto.baccini@unisi.it



Research Evaluation, 26(4), 2017, 353-357
doi: 10.1093/reseval/rvx013
Advance Access Publication Date: 27 April 2017
Article

OXFORD

A letter on Ancaiani et al. 'Evaluating scientific research in Italy: the 2004-10 research evaluation exercise'

Alberto Baccini[1] and Giuseppe De Nicolao[2]

This letter documents some problems in Ancaiani et al. (2015). Namely the evaluation of concord-
1 ance, based on Cohen's kappa, reported by Ancaiani et al. was not computed on the whole ran-
dom sample of 9,199 articles, but on a subset of 7,597 articles. The kappas relative to the whole
random sample were in the range 0.07–0.15, indicating an unacceptable agreement between peer
2 review and bibliometrics. The subset was obtained by non-random exclusion of all articles for
which bibliometrics produced an uncertain classification; these raw data were not disclosed, so
3 that concordance analysis is not reproducible. The VQR-weighted kappa for Area 13 reported by
Ancaiani et al. is higher than that reported by Area 13 panel and confirmed by Bertocchi et al.
4 (2015), a difference explained by the use, under the same name, of two different set of weights.
Two values of kappa reported by Ancaiani et al. differ from the corresponding ones published in
5 the official report. Results reported by Ancaiani et al. do not support a good concordance between
peer review and bibliometrics. As a consequence, the use of both techniques introduced systematic distortions in the final results of the Italian research assessment exercise. The conclusion that it is possible to use both technique as interchangeable in a research assessment exercise appears to be unsound, by being based on a misinterpretation of the statistical significance of kappa values.

Protocollo 5X5 vs. protocollo 4X4

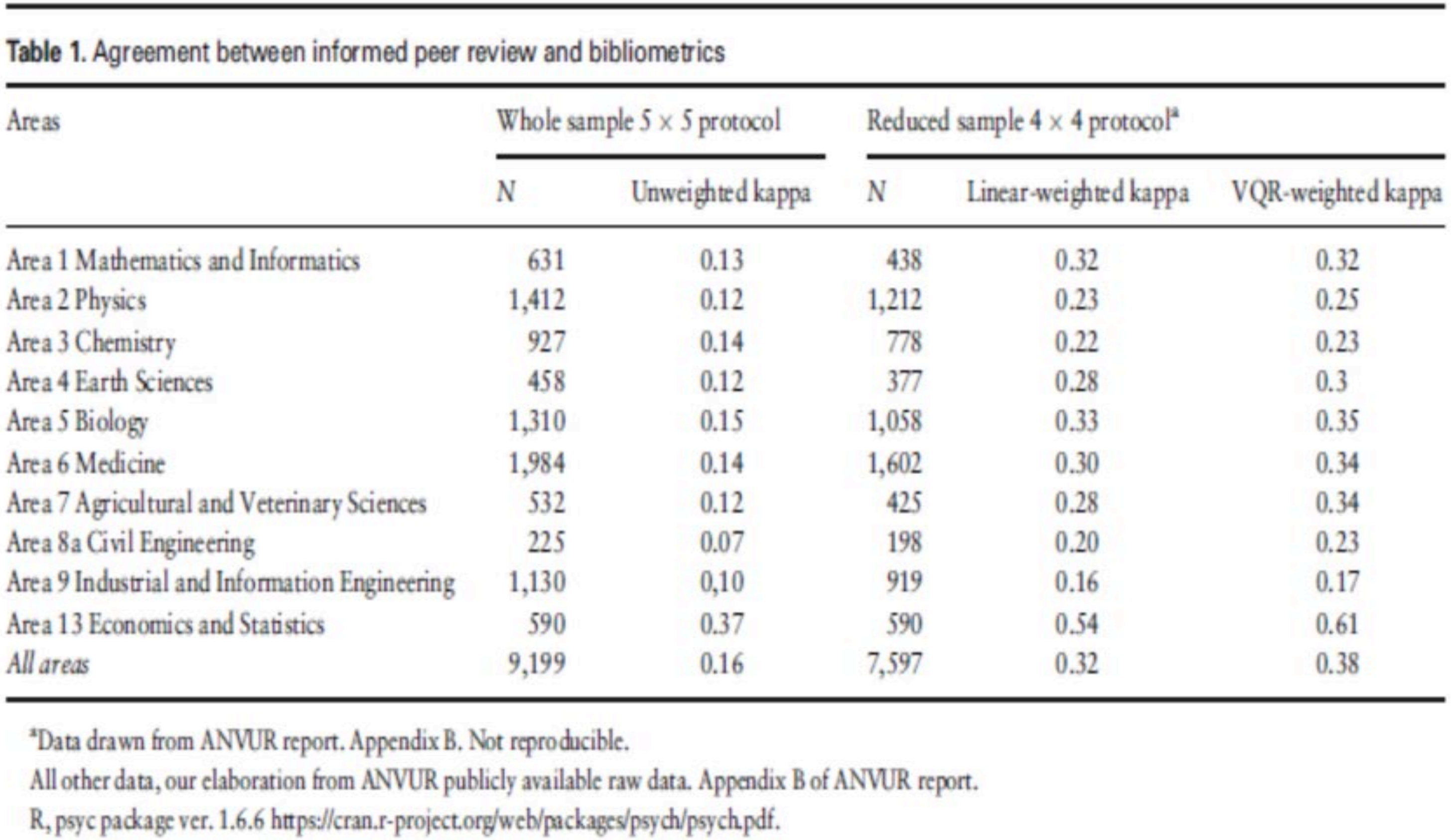


Table 1. Agreement between informed peer review and bibliometrics

Areas	Whole sample 5 × 5 protocol		Reduced sample 4 × 4 protocol[a]		
	N	Unweighted kappa	N	Linear-weighted kappa	VQR-weighted kappa
Area 1 Mathematics and Informatics	631	0.13	438	0.32	0.32
Area 2 Physics	1,412	0.12	1,212	0.23	0.25
Area 3 Chemistry	927	0.14	778	0.22	0.23
Area 4 Earth Sciences	458	0.12	377	0.28	0.3
Area 5 Biology	1,310	0.15	1,058	0.33	0.35
Area 6 Medicine	1,984	0.14	1,602	0.30	0.34
Area 7 Agricultural and Veterinary Sciences	532	0.12	425	0.28	0.34
Area 8a Civil Engineering	225	0.07	198	0.20	0.23
Area 9 Industrial and Information Engineering	1,130	0,10	919	0.16	0.17
Area 13 Economics and Statistics	590	0.37	590	0.54	0.61
All areas	9,199	0.16	7,597	0.32	0.38

[a] Data drawn from ANVUR report. Appendix B. Not reproducible.

All other data, our elaboration from ANVUR publicly available raw data. Appendix B of ANVUR report.

R, psyc package ver. 1.6.6 https://cran.r-project.org/web/packages/psych/psych.pdf.

Protocollo 5X5 vs. protocollo 4X4

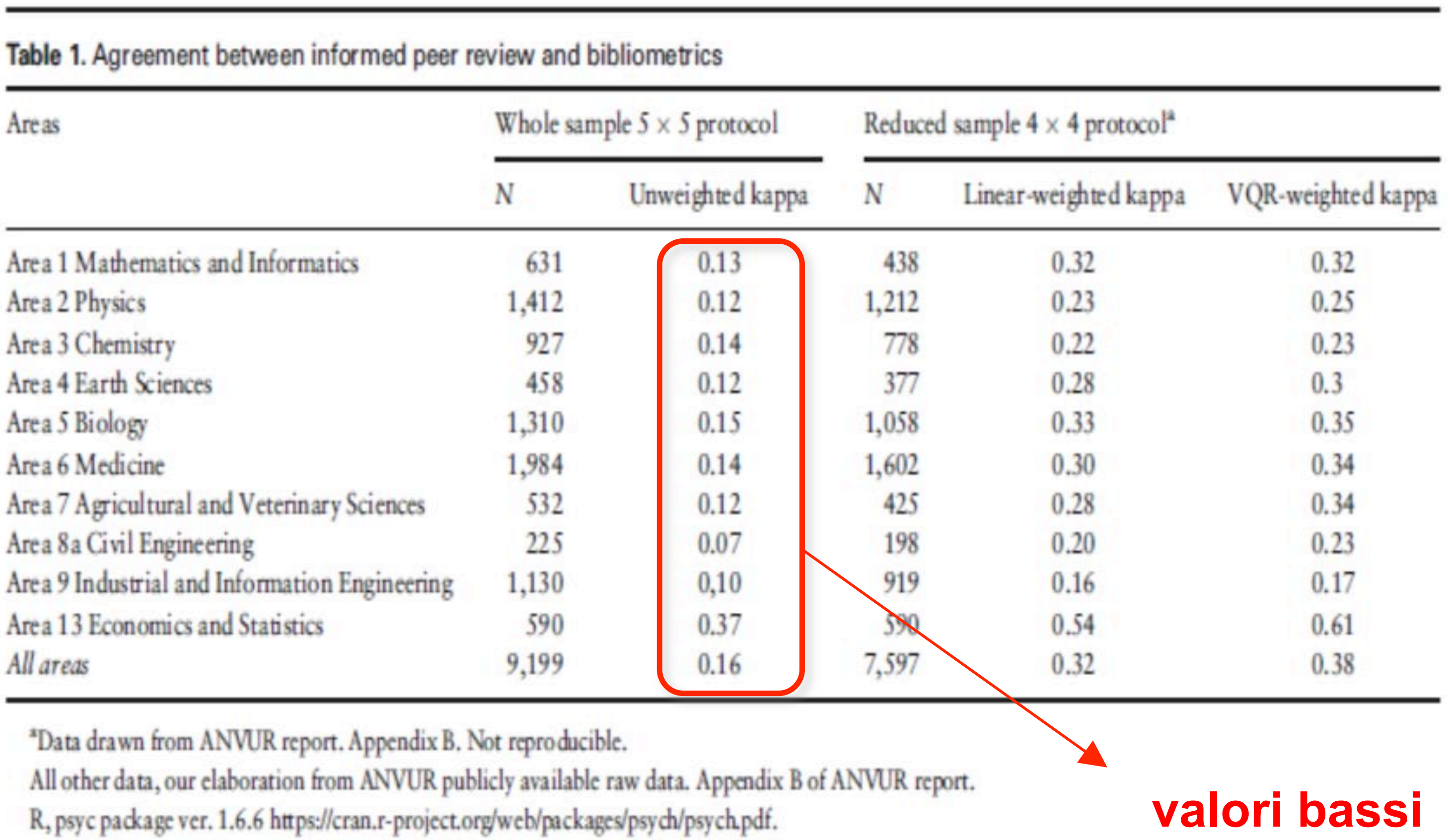


Table 1. Agreement between informed peer review and bibliometrics

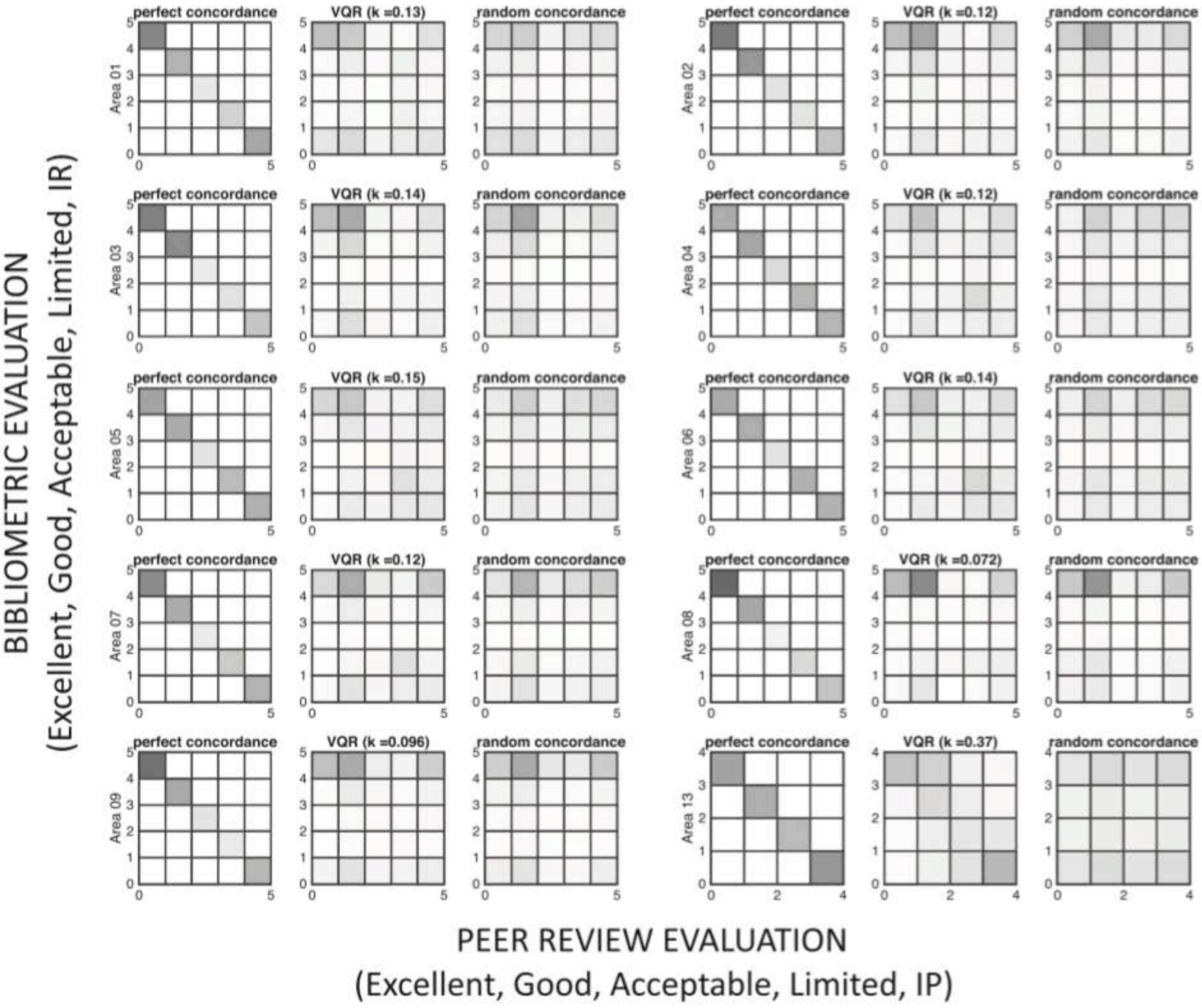
Areas	Whole sample 5 × 5 protocol		Reduced sample 4 × 4 protocol[a]		
	N	Unweighted kappa	N	Linear-weighted kappa	VQR-weighted kappa
Area 1 Mathematics and Informatics	631	0.13	438	0.32	0.32
Area 2 Physics	1,412	0.12	1,212	0.23	0.25
Area 3 Chemistry	927	0.14	778	0.22	0.23
Area 4 Earth Sciences	458	0.12	377	0.28	0.3
Area 5 Biology	1,310	0.15	1,058	0.33	0.35
Area 6 Medicine	1,984	0.14	1,602	0.30	0.34
Area 7 Agricultural and Veterinary Sciences	532	0.12	425	0.28	0.34
Area 8a Civil Engineering	225	0.07	198	0.20	0.23
Area 9 Industrial and Information Engineering	1,130	0,10	919	0.16	0.17
Area 13 Economics and Statistics	590	0.37	590	0.54	0.61
All areas	9,199	0.16	7,597	0.32	0.38

[a] Data drawn from ANVUR report. Appendix B. Not reproducible.

All other data, our elaboration from ANVUR publicly available raw data. Appendix B of ANVUR report.

R, psyc package ver. 1.6.6 https://cran.r-project.org/web/packages/psych/psych.pdf.

valori bassi di kappa non pubblicati da ANVUR


perfect concordance
VQR (k =0.13)
random concordance
perfect concordance
VQR (k =0.12)
random concordance
Area 01
Area 02
perfect concordance
VQR (k =0.14)
random concordance
perfect concordance
VQR (k =0.12)
random concordance
Area 03
Area 04
perfect concordance
VQR (k =0.15)
random concordance
perfect concordance
VQR (k =0.14)
random concordance
Area 05
Area 06
perfect concordance
VQR (k =0.12)
random concordance
perfect concordance
VQR (k =0.072)
random concordance
Area 07
Area 08
perfect concordance
VQR (k =0.096)
random concordance
perfect concordance
VQR (k =0.37)
random concordance
Area 09
Area 13
0
1
2
3
4
5
BIBLIOMETRIC EVALUATION
(Excellent, Good, Acceptable, Limited, IR)
PEER REVIEW EVALUATION
(Excellent, Good, Acceptable, Limited, IP)

Ancaiani et al. 2015

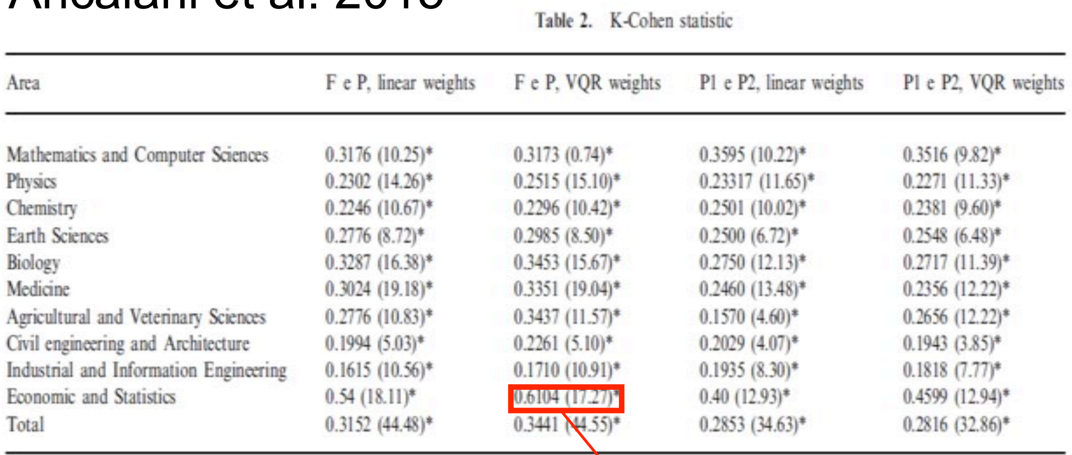
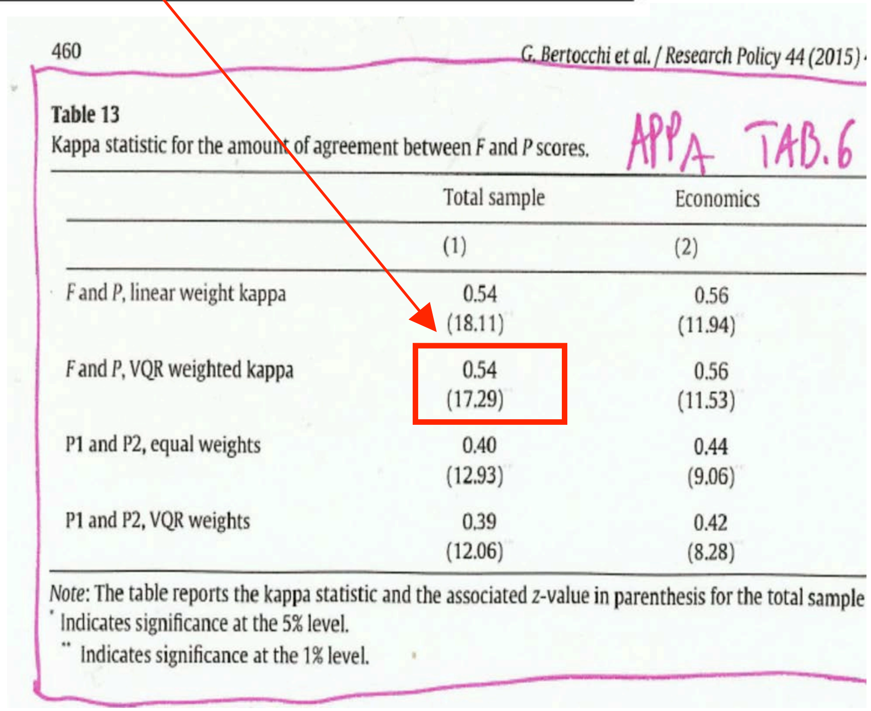


Table 2. K-Cohen statistic

Area	F e P, linear weights	F e P, VQR weights	P1 e P2, linear weights	P1 e P2, VQR weights
Mathematics and Computer Sciences	0.3176 (10.25)*	0.3173 (0.74)*	0.3595 (10.22)*	0.3516 (9.82)*
Physics	0.2302 (14.26)*	0.2515 (15.10)*	0.23317 (11.65)*	0.2271 (11.33)*
Chemistry	0.2246 (10.67)*	0.2296 (10.42)*	0.2501 (10.02)*	0.2381 (9.60)*
Earth Sciences	0.2776 (8.72)*	0.2985 (8.50)*	0.2500 (6.72)*	0.2548 (6.48)*
Biology	0.3287 (16.38)*	0.3453 (15.67)*	0.2750 (12.13)*	0.2717 (11.39)*
Medicine	0.3024 (19.18)*	0.3351 (19.04)*	0.2460 (13.48)*	0.2356 (12.22)*
Agricultural and Veterinary Sciences	0.2776 (10.83)*	0.3437 (11.57)*	0.1570 (4.60)*	0.2656 (12.22)*
Civil engineering and Architecture	0.1994 (5.03)*	0.2261 (5.10)*	0.2029 (4.07)*	0.1943 (3.85)*
Industrial and Information Engineering	0.1615 (10.56)*	0.1710 (10.91)*	0.1935 (8.30)*	0.1818 (7.77)*
Economic and Statistics	0.54 (18.11)*	0.6104 (17.27)*	0.40 (12.93)*	0.4599 (12.94)*
Total	0.3152 (44.48)*	0.3441 (44.55)*	0.2853 (34.63)*	0.2816 (32.86)*



460 G. Bertocchi et al. / Research Policy 44 (2015)

Table 13
Kappa statistic for the amount of agreement between *F* and *P* scores.

APPA TAB.6

	Total sample	Economics
	(1)	(2)
F and *P*, linear weight kappa	0.54 (18.11)	0.56 (11.94)
F and *P*, VQR weighted kappa	0.54 (17.29)	0.56 (11.53)
P1 and P2, equal weights	0.40 (12.93)	0.44 (9.06)
P1 and P2, VQR weights	0.39 (12.06)	0.42 (8.28)

Note: The table reports the kappa statistic and the associated z-value in parenthesis for the total sample

* Indicates significance at the 5% level.

** Indicates significance at the 1% level.

Errore nei dati o altro?

Altro: ci sono due sistemi di pesi chiamati nello stesso modo

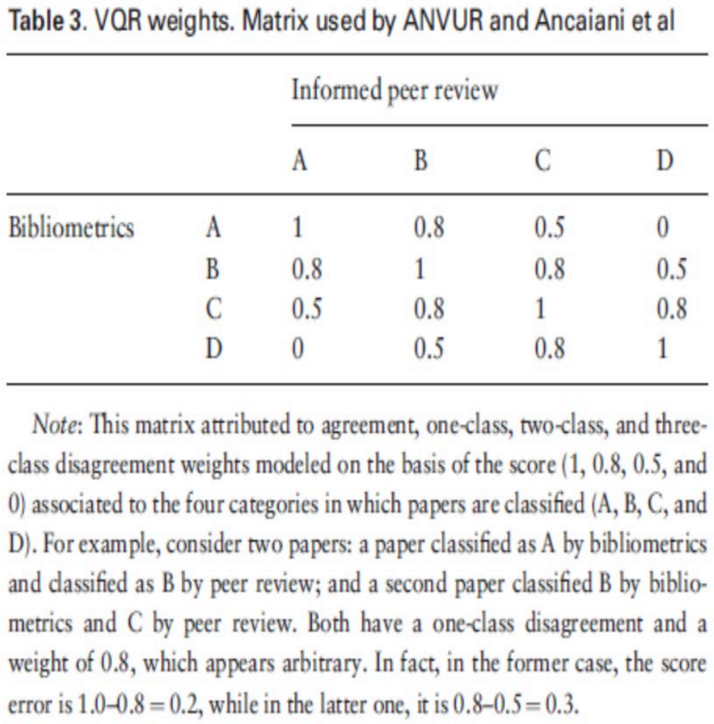


Table 3. VQR weights. Matrix used by ANVUR and Ancaiani et al

		Informed peer review			
		A	B	C	D
Bibliometrics	A	1	0.8	0.5	0
	B	0.8	1	0.8	0.5
	C	0.5	0.8	1	0.8
	D	0	0.5	0.8	1

Note: This matrix attributed to agreement, one-class, two-class, and three-class disagreement weights modeled on the basis of the score (1, 0.8, 0.5, and 0) associated to the four categories in which papers are classified (A, B, C, and D). For example, consider two papers: a paper classified as A by bibliometrics and classified as B by peer review; and a second paper classified B by bibliometrics and C by peer review. Both have a one-class disagreement and a weight of 0.8, which appears arbitrary. In fact, in the former case, the score error is 1.0–0.8 = 0.2, while in the latter one, it is 0.8–0.5 = 0.3.

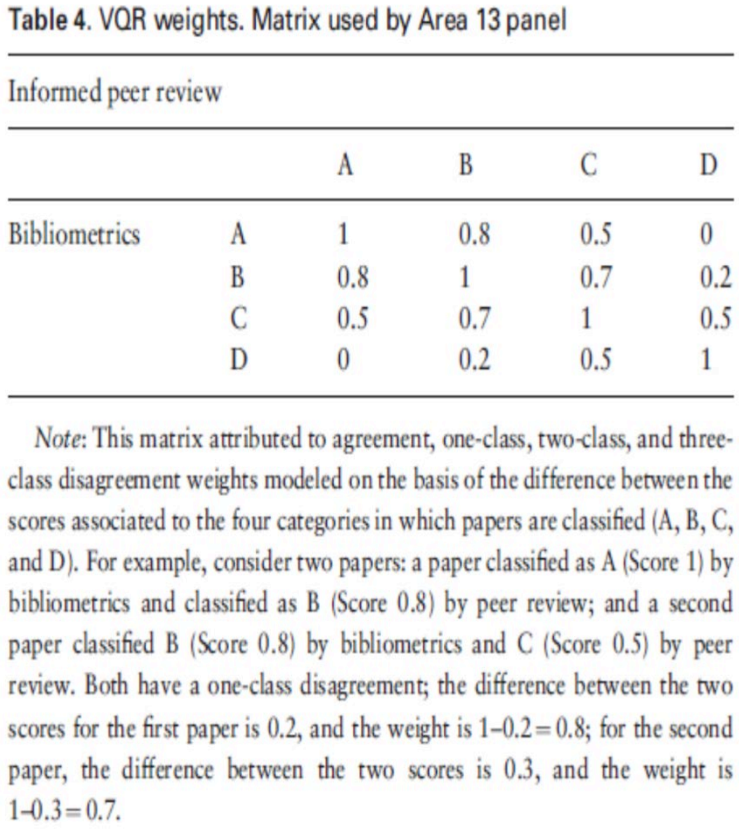
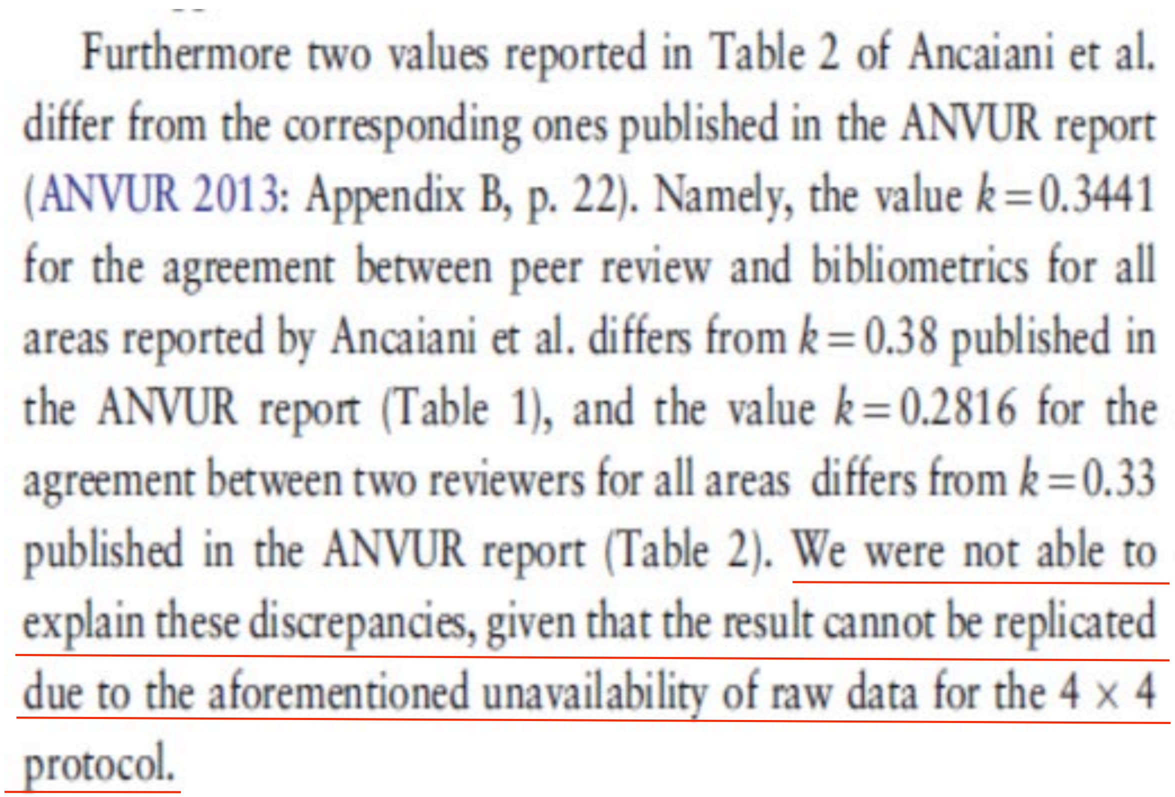


Table 4. VQR weights. Matrix used by Area 13 panel

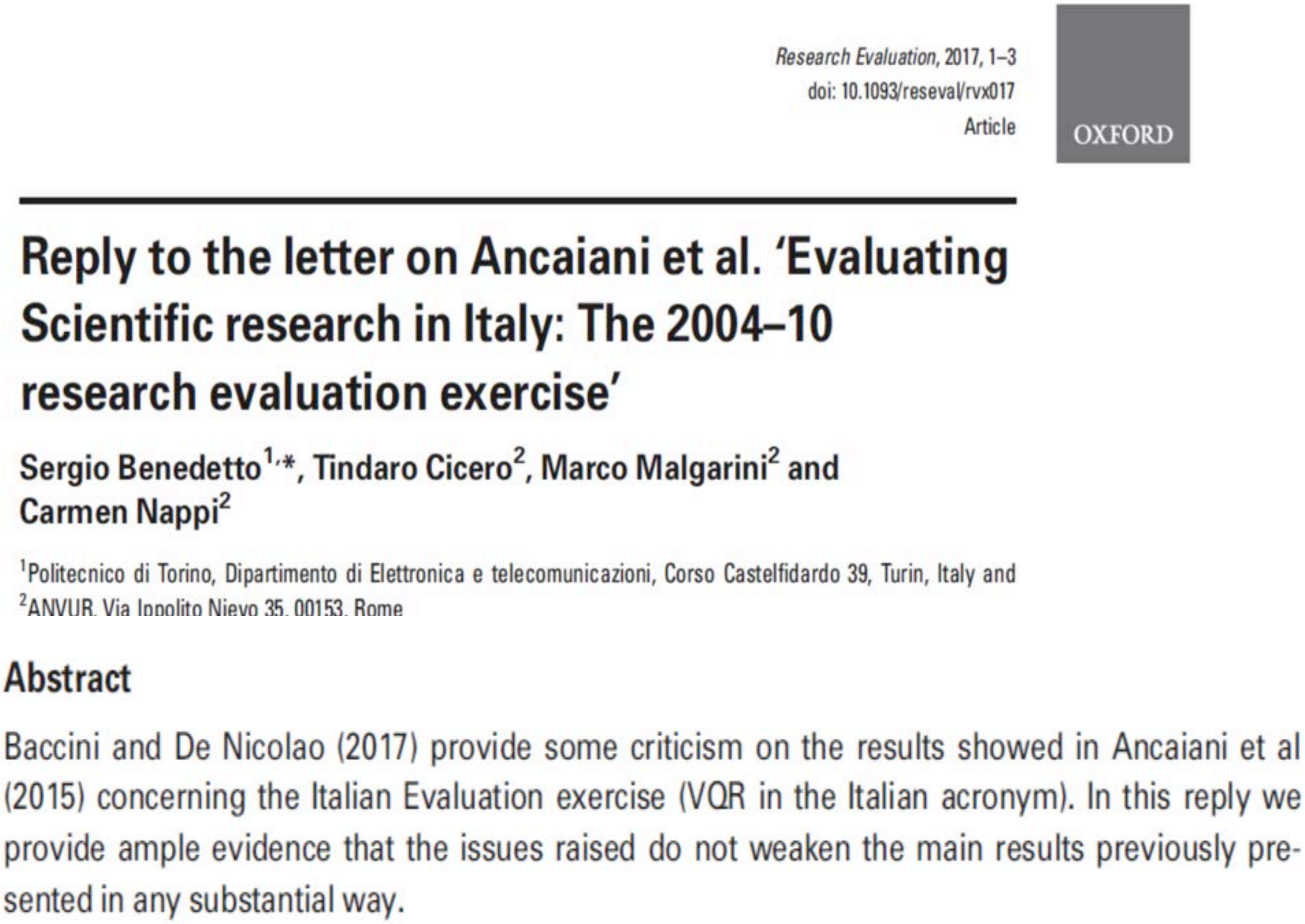
Informed peer review					
		A	B	C	D
Bibliometrics	A	1	0.8	0.5	0
	B	0.8	1	0.7	0.2
	C	0.5	0.7	1	0.5
	D	0	0.2	0.5	1

Note: This matrix attributed to agreement, one-class, two-class, and three-class disagreement weights modeled on the basis of the difference between the scores associated to the four categories in which papers are classified (A, B, C, and D). For example, consider two papers: a paper classified as A (Score 1) by bibliometrics and classified as B (Score 0.8) by peer review; and a second paper classified B (Score 0.8) by bibliometrics and C (Score 0.5) by peer review. Both have a one-class disagreement; the difference between the two scores for the first paper is 0.2, and the weight is 1–0.2 = 0.8; for the second paper, the difference between the two scores is 0.3, and the weight is 1–0.3 = 0.7.

Altri dati che non quadrano. Perché?



Furthermore two values reported in Table 2 of Ancaiani et al. differ from the corresponding ones published in the ANVUR report (ANVUR 2013: Appendix B, p. 22). Namely, the value $k = 0.3441$ for the agreement between peer review and bibliometrics for all areas reported by Ancaiani et al. differs from $k = 0.38$ published in the ANVUR report (Table 1), and the value $k = 0.2816$ for the agreement between two reviewers for all areas differs from $k = 0.33$ published in the ANVUR report (Table 2). We were not able to explain these discrepancies, given that the result cannot be replicated due to the aforementioned unavailability of raw data for the 4 × 4 protocol.



Research Evaluation, 2017, 1–3
doi: 10.1093/reseval/rvx017
Article

Reply to the letter on Ancaiani et al. 'Evaluating Scientific research in Italy: The 2004–10 research evaluation exercise'

Sergio Benedetto[1,*], Tindaro Cicero[2], Marco Malgarini[2] and Carmen Nappi[2]

[1]Politecnico di Torino, Dipartimento di Elettronica e telecomunicazioni, Corso Castelfidardo 39, Turin, Italy and [2]ANVUR, Via Ippolito Nievo 35, 00153, Rome

Abstract

Baccini and De Nicolao (2017) provide some criticism on the results showed in Ancaiani et al (2015) concerning the Italian Evaluation exercise (VQR in the Italian acronym). In this reply we provide ample evidence that the issues raised do not weaken the main results previously presented in any substantial way.

Research Policy & Evaluation

A journal on

RT

Open access, peer-reviewed journal.

RT. A Journal on Research Policy & Evaluation **1** (2017)
Submitted on 10 July 2017, published on 22 July 2017, open for comments

Doi: 10.13130/2282-5398/8872

Errors and secret data in the Italian research assessment exercise. A comment to a reply

Alberto Baccini[*], Giuseppe De Nicolao[**]

Errori inspiegabili nella replica

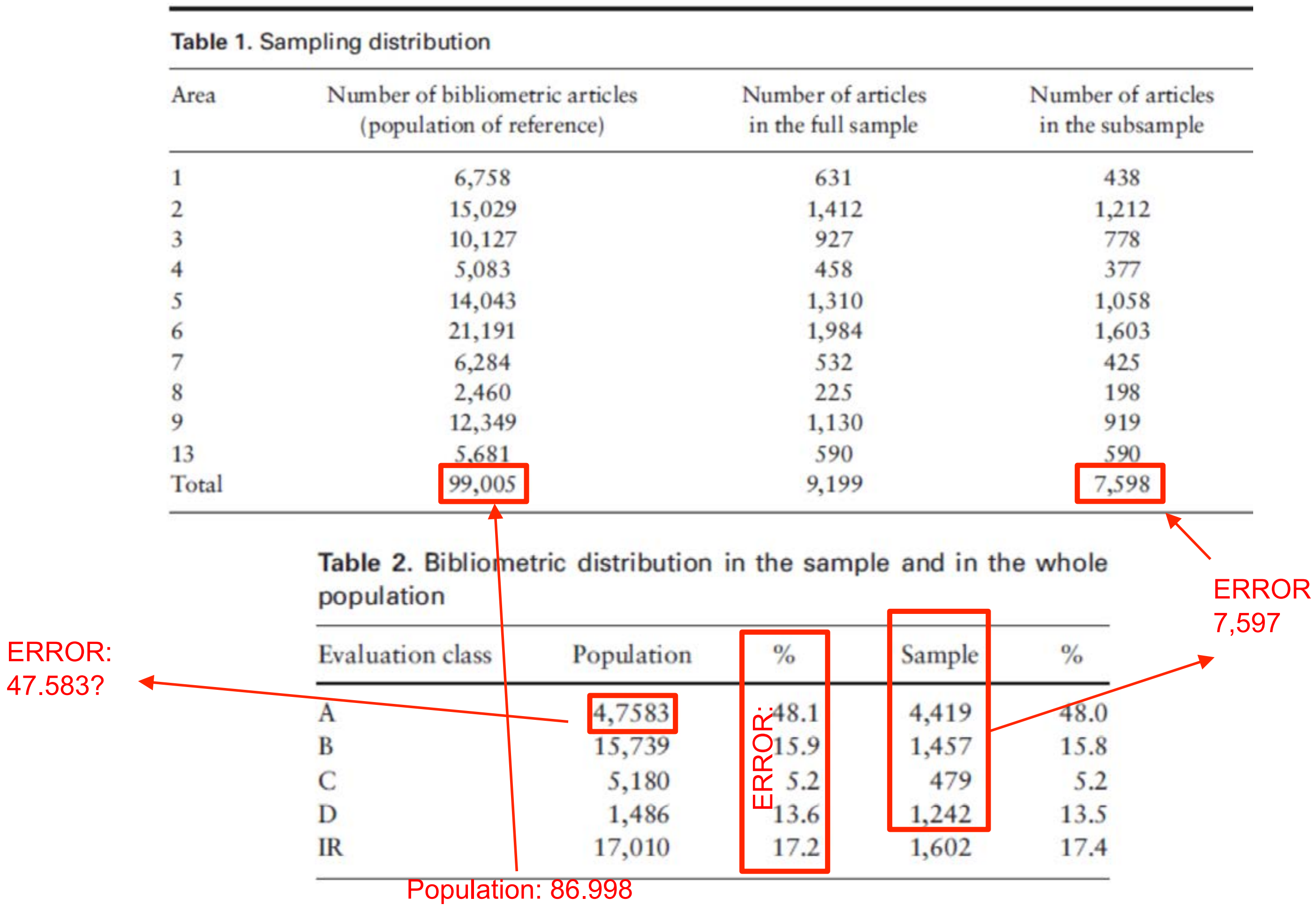


Table 1. Sampling distribution

Area	Number of bibliometric articles (population of reference)	Number of articles in the full sample	Number of articles in the subsample
1	6,758	631	438
2	15,029	1,412	1,212
3	10,127	927	778
4	5,083	458	377
5	14,043	1,310	1,058
6	21,191	1,984	1,603
7	6,284	532	425
8	2,460	225	198
9	12,349	1,130	919
13	5,681	590	590
Total	99,005	9,199	7,598

Table 2. Bibliometric distribution in the sample and in the whole population

Evaluation class	Population	%	Sample	%
A	4,7583	48.1	4,419	48.0
B	15,739	15.9	1,457	15.8
C	5,180	5.2	479	5.2
D	1,486	13.6	1,242	13.5
IR	17,010	17.2	1,602	17.4

ERROR: 47.583?

ERROR:

Population: 86.998

ERROR 7,597

7. Conclusioni

ANVUR e la giustificazione della politica italiana per la ricerca

Why this extraordinary dissemination effort was produced by scholars working for ANVUR?

Probably because the publication in scholarly journals represent an ex-post justification of the unprecedented dual system of evaluation developed and applied by ANVUR.

The metodology and results of the research assessment are justified ex-post by papers written by scholars that have developed and applied the methodology adopted by the Italian government.

Moreover, the results of these papers cannot be replicated because the data were not made available to scholars other than those working for ANVUR.

Politica vaccinale

Government prescribes a new mandatory vaccine in compliance with the recommendation of a report issued by an agency such as the Food and Drug Administration.

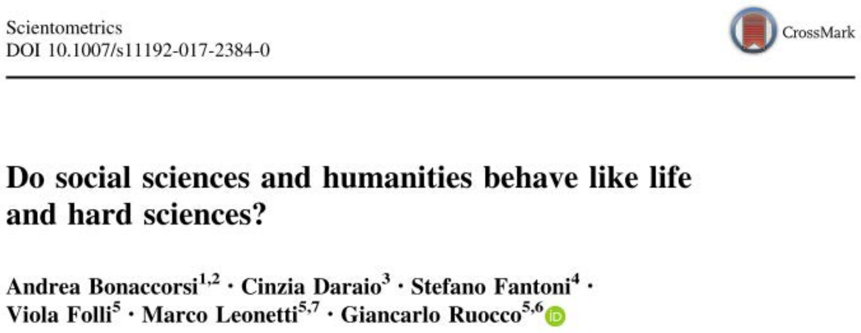
A couple of years after the mandatory adoption, scholarly journals publish articles, authored by members of the FDA committee that issued the report.

Although not declared, these articles reproduce contents and conclusions of the FDA report, thus providing a *de facto* – though *ex post* - scientific justification of the report itself.

When independent scholars ask data for replicating results, the agency does not reply or, alternatively, denies the data alleging that they are confidential.

Fortunately, this is not the way health decisions are usually taken.

Inquinamento della letteratura

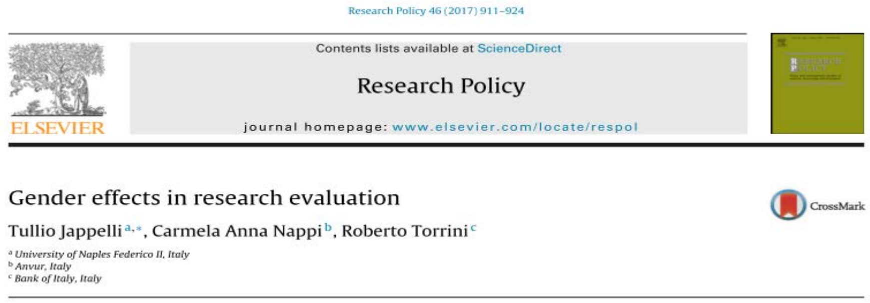


Scientometrics
DOI 10.1007/s11192-017-2384-0

CrossMark

Do social sciences and humanities behave like life and hard sciences?

Andrea Bonaccorsi[1,2] · Cinzia Daraio[3] · Stefano Fantoni[4] · Viola Folli[5] · Marco Leonetti[5,7] · Giancarlo Ruocco[5,6]



Research Policy 46 (2017) 911–924

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

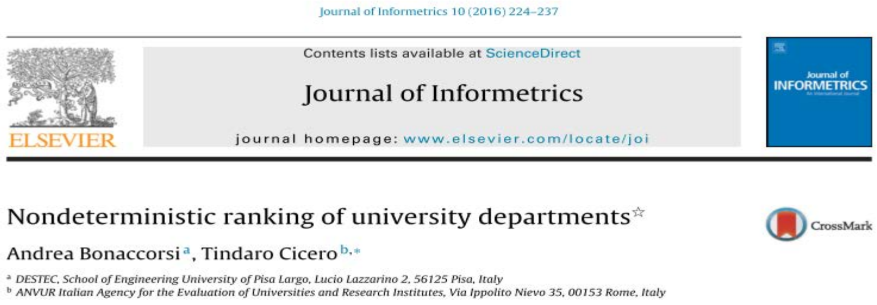
ELSEVIER

Gender effects in research evaluation

CrossMark

Tullio Jappelli[a,*], Carmela Anna Nappi[b], Roberto Torrini[c]

[a] University of Naples Federico II, Italy
[b] Anvur, Italy
[c] Bank of Italy, Italy



Journal of Informetrics 10 (2016) 224–237

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

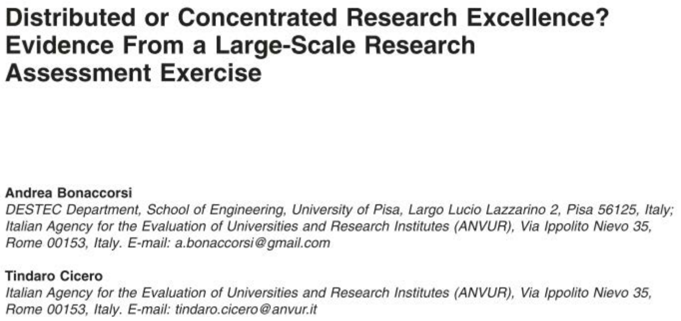
ELSEVIER

Nondeterministic ranking of university departments☆

CrossMark

Andrea Bonaccorsi[a], Tindaro Cicero[b,*]

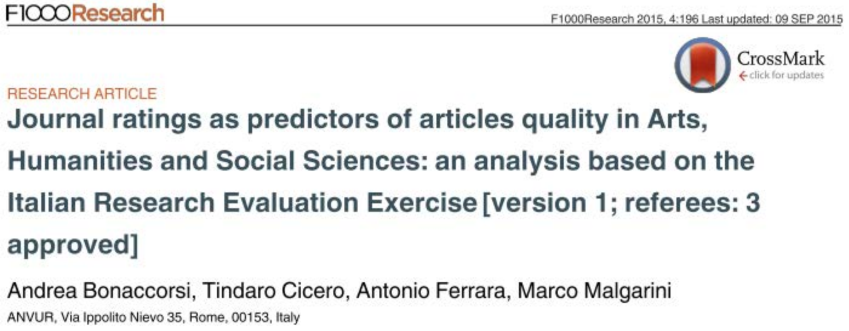
[a] DESTEC, School of Engineering University of Pisa Largo, Lucio Lazzarino 2, 56125 Pisa, Italy
[b] ANVUR Italian Agency for the Evaluation of Universities and Research Institutes, Via Ippolito Nievo 35, 00153 Rome, Italy



Distributed or Concentrated Research Excellence? Evidence From a Large-Scale Research Assessment Exercise

Andrea Bonaccorsi
DESTEC Department, School of Engineering, University of Pisa, Largo Lucio Lazzarino 2, Pisa 56125, Italy; Italian Agency for the Evaluation of Universities and Research Institutes (ANVUR), Via Ippolito Nievo 35, Rome 00153, Italy. E-mail: a.bonaccorsi@gmail.com

Tindaro Cicero
Italian Agency for the Evaluation of Universities and Research Institutes (ANVUR), Via Ippolito Nievo 35, Rome 00153, Italy. E-mail: tindaro.cicero@anvur.it



F1000Research

F1000Research 2015, 4:196 Last updated: 09 SEP 2015

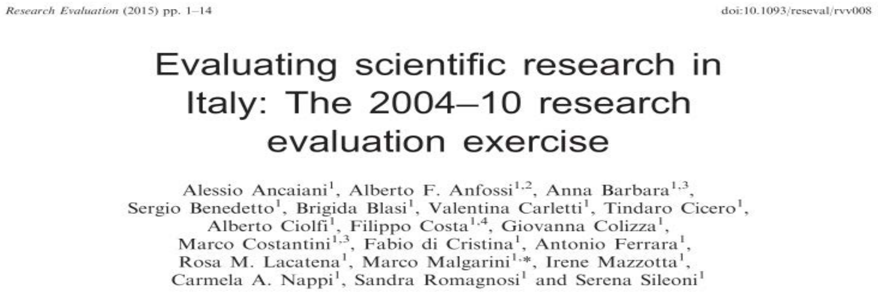
CrossMark
click for updates

RESEARCH ARTICLE

Journal ratings as predictors of articles quality in Arts, Humanities and Social Sciences: an analysis based on the Italian Research Evaluation Exercise [version 1; referees: 3 approved]

Andrea Bonaccorsi, Tindaro Cicero, Antonio Ferrara, Marco Malgarini

ANVUR, Via Ippolito Nievo 35, Rome, 00153, Italy



Research Evaluation (2015) pp. 1–14 doi:10.1093/reseval/rvv008

Evaluating scientific research in Italy: The 2004–10 research evaluation exercise

Alessio Ancaiani[1], Alberto F. Anfossi[1,2], Anna Barbara[1,3], Sergio Benedetto[1], Brigida Blasi[1], Valentina Carletti[1], Tindaro Cicero[1], Alberto Ciolfi[1], Filippo Costa[1,4], Giovanna Colizza[1], Marco Costantini[1,3], Fabio di Cristina[1], Antonio Ferrara[1], Rosa M. Lacatena[1], Marco Malgarini[1,*], Irene Mazzotta[1], Carmela A. Nappi[1], Sandra Romagnosi[1] and Serena Sileoni[1]