Scholarly Assessment Reports

Bonaccorsi, A. (2020). Two Decades of Research Assessment in Italy. Addressing the Criticisms. *Scholarly Assessment Reports*, 2(1): 17. DOI: https://doi.org/10.29024/sar.28

RESEARCH

Two Decades of Research Assessment in Italy. Addressing the Criticisms

Andrea Bonaccorsi
University of Pisa, IT
a.bonaccorsi@gmail.com

Italy is the single largest country in Continental Europe to have adopted a regular and mandatory research assessment approach, involving all researchers at universities and Public Research Organizations (PROs), with impact on performance-based funding. With more than 180,000 products, evaluated by more than 14,000 referees, the 2004–2010 exercise carried out by a newly created Agency (ANVUR) was one of the largest ever carried out. It has adopted a peculiar mixed-methodology approach, using peer review in Social Sciences and Humanities (SSH) and bibliometrics in STEM disciplines. The approach has raised a number of conceptual and technical issues. In parallel a major reform of academic recruitment has introduced quantitative indicators as threshold values for candidates to the National Scientific Habilitation. This procedure has been made possible by a massive exercise of classification and rating of journals. The paper addresses the most important criticisms raised against these research assessment initiatives and checks their arguments against empirical evidence. The paper also addresses the controversial issue of unintended and negative consequences of research assessment. The final section offers some policy highlights.

Keywords: criticism to research assessment; unintended consequence; Impact factor; Peer review; journal rating

1. Introduction

In a companion paper (Bonaccorsi, 2020) I have described the Italian experience of research assessment in the last two decades (2000–2020). Research assessment in Italy has been similar in size to the RAE/REF in the UK, but has been the most comprehensive by scope and impact. It has included a formal assessment exercise (VQR), a journal rating exercise, and the production of quantitative indicators used for the National Scientific Habilitation. In a relatively short time span of a few years, it has affected the funding and reputation of universities, the internal allocation of resources across disciplines and departments, and the careers of researchers. Given the pervasiveness and impact, it is not surprising that it has attracted lot of criticism.

In this paper I will try to address systematically the most important criticisms that have been published in refereed journals. I will deliberatly ignore the national press (including newspapers) and the web media, with a few exceptions only. I will mainly use papers in English, for the benefit of readers. For a rich account of the debate in Italian language see the introduction in Fassari and Valentini (2020), with more than 200 references. I hope to be able to give full justice to the authors and to the counterarguments offered by ANVUR.

In reviewing the criticism I will take into account an important document produced by an independent expert panel upon request of ANVUR (Group of experts, 2019).[1] This report includes a number of recommendations.

The discussion will be somewhat technical, that is, related mainly to methodology, choice of indicators, choice of algorithms and aggregation formulas. Readers more interested in high level discussions about the purposes, objectives and impact of research assessment might benefit from reading my companion paper before this one.

[1] The Expert panel was formed by Claudio Galderisi (HCERES and Université de Poitiers), Mauro Perretti (Chair, Queen Mary University of London), Nuria Sebastian Galles (Universitat Pompeu Fabra, Barcelona), Thed van Leeuwen (Leiden University). I examined their CVs from publicly available sources. The members of the panel are distinguished scientists affiliated to high level institutions in various disciplines, such as humanities (Galderisi), medicine (Perretti) and neuroscience (Sebastian Galles), while van Leeuwen is among the leading authors in bibliometrics.

2. Use of bibliometrics and peer review

ANVUR decided to adopt not just one methodology, but two, i.e. bibliometrics and peer review. This decision followed the policy mandate of the law creating the Agency, which considered both methodologies acceptable (see my companion paper for the policy discussion). It also followed the international state of the art, which recommends the adoption of methodologies that reflect the differences between scientific fields, as witnessed by the San Francisco Declaration on Research Assessment (DORA, 2012), the IEEE statement on bibliometrics (2013), the Leiden Manifesto (Hicks et al. 2015), and the position paper by the COST initiative ENRESSH (Ochsner et al. 2020). In addition this decision was made following a large literature on the evaluation of research in Social Sciences and Humanities (SSH), according to which citation analysis is not appropriate to evaluate research in these fields (Moed, 2005) and bibliometric databases do not offer adequate coverage of the scientific production. In short, bibliometrics cannot be used in SSH (Moed et al. 2002; Hicks, 2004; Archambault et al. 2006; Nederhof, 2006). Given the intense use of books as communication channel, the only acceptable methodology is peer review (Larivière et al. 2006; Giménez-Toledo, 2020). The later debate in the literature on research assessment largely confirmed this assumption (Ochsner, Hug and Daniel, 2016; Bonaccorsi et al. 2017; Bonaccorsi, 2018a). At the same time, extending peer review to all fields was not feasible, given the budget constraint and the need to mobilize a very large number of referees in order to avoid the overload.

There are two different problems here. The first is that when aggregating at the level of universities the scores from VQR there might be a *composition effect* depending on the share of subjects evaluated using peer review or bibliometrics.

The second issue is that a certain share of articles was sent to peer review and not evaluated bibliometrically (in particular, very recent papers) also in Science, Technology, Engineering and Mathematics (STEM) panels and the peer review scores were then aggregated with those obtained with bibliometric indicators. In particular papers were sent to peer review when they were too young (citations were not considered adequate to represent quality); the topics were new, emerging, or interdisciplinary; the institutions requested to send the paper to peer review and the request was accepted by the evaluation panel (*Gruppo di Esperti della Valutazione,* GEV) and, finally, when bibliometric indicators were in sharp conflict. If there is no or a poor correlation between the scores obtained with peer review and with bibliometrics, then the composition and aggregation effects may introduce distorsions. Therefore it becomes crucial to examine the correlation.

On the first issue some authors criticized the joint use of different methodologies. According to Abramo, D'Angelo and Costa al. (2011) it would be better to use only bibliometrics, and in particular, to submit all products (not a sample) to evaluation (see below for the second part of the argument). Surprisingly, these authors tell nothing about the evaluation of research for those fields in which indexed journals do not represent the main scientific production, i.e. SSH, in particular Humanities. If I am not wrong, I have found no mention at all of this issue in papers criticizing the use of dual methodologies.

As a matter of fact the individual scores are normalized on the basis of the national average by discipline, so that the composition effect is sterilized. After the first VQR ANVUR addressed the issue of compositional effects by building a virtual aggregate score, using a procedure initially suggested by a physicist at the University of Florence, Prof. Poggi (Poggi, 2014; Poggi e Nappi, 2014). These calculations were used by the Ministry to allocate funding to excellent departments after VQR 2011–2014.[2] The procedure is based on the calculation of virtual scores that wipe out the variability in the composition of departments and universities by discipline. Thus the combination of bibliometrics in STEM disciplines and peer review in SSH does not imply distortions at the level of universities.

On the second issue, a more technical objection was that bibliometrics and peer review have different properties in terms of the distribution of scores for the same field (Baccini, 2016; Baccini and De Nicolao, 2016). This effect was noted immediately by ANVUR in the first VQR Report and subsequently elaborated by Ancaiani et al. (2015). As a matter of fact, when both sources of evaluation are available, it turns out that the average score in peer review is lower than in bibliometrics. This issue has generated another stream of controversies. In the expert panel in Economics and Statistics, a random sample of articles already evaluated with bibliometric indicators was also sent to peer review. Bertocchi et al. (2015) report that the Cohen's kappa coefficient that measure the agreement between the two types of evaluation is in the range 0.30–0.50, or moderate to fair. Baccini and De Nicolao (2016; 2017) submitted the same data to another test and concluded that the degree of agreement was much lower. Consequently, the allocation of articles to peer review or bibliometrics may result into distortions.

My interpretation of the debate is as follows. The comparison between peer review and bibliometrics is a classical issue in the research assessment literature, in particular following the RAE and REF experiences in the UK. Several authors examined the correlation between the scores of various collections of articles. I find that there has been some misconception in this methodological discussion.

From a theoretical point of view I see research evaluation as a practical judgment, in which expert persons try to summarize a vast amount of knowledge related to a specific piece of research according to several criteria. I do not expect there is systematic agreement on all individual pieces of research. To be honest, the agreement between two referees is usually lower than the agreement between a single referee and any bibliometric indicator. Evaluation will always be imperfect. What we must check is whether the agreement among various sources of evaluation is larger than then it would be when choosing alternative combinations of sources of judgment.

[2] Complete documentation available at https://www.anvur.it/attivita/dipartimenti/materiali-di-approfondimento/.

For any single piece of research it is very unlikely that we have found the optimal combination. Facing with individual items to be evaluated it is common experience that the agrement among referees is very large at the opposite tails of the distribution of evaluations (i.e. for extremely brilliant and for poor products), while it is much lower in the centre of the distribution. Consequently we cannot expect strong correlation between scores assigned by referees and bibliometric indicators at the level of individual publications. In this perspective, I find too ambitious the statement, put forward by Bertocchi et al. (2015) according to which peer review and bibliometrics can be used *interchangeably*. This would require a high level of Cohen's kappa coefficient. But we do not need this level of agreement. What we need is to be sure that the error we make by using both methodologies for reasons of necessity, is acceptable *on aggregate* with respect to the ideal situation in which the products were evaluated entirely with just one of the methodologies. What we mean by acceptable must be defined ex ante. This point has been made in a compelling way by Traag and Waltman (2019) who review the large literature on peer review-bibliometric agreement.

This perspective is taken with respect to the Italian experience by a very recent paper by Traag, Malgarini and Sarlo (2020). They exploit the fact that 10% of all journal articles submitted to the VQR 2011–2014 have been also sent for peer review to two independent referees. They found that the degree of agreement between bibliometric and peer review is actually larger than the degree of agreement between two referees. This holds for citation indicators as well as for journal-level indicators. At the aggregate level (not individual level) the error introduced by combining peer review and bibliometrics is therefore not larger than it would be when using any other known combinations of methods. I find this argument appropriate.

3. Use of journal indicator

As described in Ancaiani et al. (2015) the expert groups that adopted the bibliometric methodology combined two indicators: the normalized number of citations and a journal-level indicator. A classical objection to the use of journal indicators in research assessment is that these indicators capture only the mean of the distribution of citations, while it is well known that citations are distributed in a very skewed way (Radner, 1998; Todorov and Glänzel, 1988; Glänzel and Moed, 2002; Leydesdorff, 2008). In addition the impact factor is not transparent (Seglen, 1997; Mingers and Leydesdorff, 2015), the time frame is too short and the correlation between impact factors and average citation impact of articles became weaker over time (Lozano et al 2012). Consequently, most authors warn against the use of journal indicators for the evaluation of individual researchers (Moed and van Leeuwen, 1996; van Leeuwen and Moed, 2005; Jarwal et al. 2009; Marx and Bornmann, 2013; for a recent and updated review see Larivière and Sugimoto, 2019). Based on these arguments, the use of journal indicators in the bibliometric evaluation of VQR has been criticised in general terms, although there is no published article addressing the potential distortions of its use.

In order to examine whether the inclusion of journal-level indicators has produced serious distortions it is important to check whether (i) it is combined with other indicators; (ii) it is used for individual evaluation.

First of all, journal indicators are used jointly with normalized citations. These two indicators are assumed as (imperfect) signals of the quality of research, as reflected in the practice of the relevant research community. The final score is constructed automatically if the signals are strongly correlated, while it is assigned by experts (peer review) if the signals show little correlation.

Second, while the unit of analysis is the individual product submitted by universities, upon initiative of individual researchers, it is clear that the unit of evaluation is not the individual researcher at all. It is an aggregate of researchers: the scientific field (*Settore Scientifico Disciplinare*, or SSD in the Italian administrative language), with a minimum of 4 researchers in order to preserve confidentiality, the department, the university or PRO.

Following the recommendations of the literature, journal-level indicators have not been used for the evaluation of individual researchers. In the procedure for the assessment of individual researchers (National Scientific Habilitation) the bibliometric indicators used by ANVUR in STEM fields do not include journal impact factors. Candidates are admitted to the qualitative evaluation by a committee of Full professors at national level if they overcome a threshold in a combination of the number of indexed publications, the count of citations, and the h-index. No mention of journal indicators at all.

Why did ANVUR include journal-level indicators, given their controversial nature? It was considered that journal indicators were useful for very recent articles, for which the citation window would be too short to formulate a fair judgment. In addition, maintaining journal impact factors as elements of the VQR evaluation was intended as a way to suggest to young researchers the importance of publishing in good international journals, rather than splitting papers in lower level ones.[3]

[3] In two articles that describe the VQR, written by ANVUR member, the methodological choice has been motivated as follows: "The journal impact, if used correctly, is considered a good proxy of the journal quality (Abramo, D'Angelo, and Di Costa 2010), being consistent with overall citation counts (Hunt et al. 2010) and with journals' rejection rates; moreover, the use of multiple metrics can reduce the risk that evaluation results are affected by editors' or authors' manipulation. Finally, the use of journal metrics may help conveying to young researchers the message that journals may differ in quality, and that it is important for the researchers to measure themselves against the most rigorous peer review procedures in order to publish articles in the most prestigious journals in any given research area" (Ancaiani et al. 2015, 245). In addition, "the use of the sole citation count may not be an appropriate indicator

Having said that I recognize that there is a large international debate on the advantages and disadvantages of using journal-based indicators. On this point it is important to keep the debate open and balance carefully the arguments, without taking strong a priori positions. A couple of recent papers reopened the debate, bringing evidence in favour of the use of journal-based indicators (Traag, 2019; Waltman and Traag, 2019).

The independent Expert group advised ANVUR to eliminate journal-level indicators from the bibliometric methodology (Expert group, 2019). The arguments are as follows.

> 3.7 The bibliometric algorithm that combines these indicators is very mechanical. The choice of databases is left to the submitter, and the relative weight of journal impact factors and article cites is estimated within database journal classifications (which do not coincide with VQR assessment clusters, and may or may not be unambiguously meaningful) and item publication date. These features enhance impartiality, but make the results somewhat opaque and hard to interpret: these mechanical criteria are quite complex, with different weights and several parameters so that the final formula may be perceived as arbitrary.

> 4.3.10 For sub-disciplines, or sub-GEVs where citation numbers are appropriate, IRAS1 could be simplified and probably optimised by using solely citation numbers, where ranking into the top percentiles will be defined objectively for each sub-discipline or sub-GEV. The use of straight citation counts thus avoids the use of Journal Impact Factors, which are a marker of the average impact of the journal and does not add to the impact of specific publications.

4. Aggregation algorithm

In STEM fields the class of merit is assigned by combining percentiles of the world distribution of citations in the same Subject Category and percentiles of journal indicators.

The combination of multiple indicators creates an obvious technical problem of aggregation. As a starting point, it would be important to state that aggregation may take place according to a variety of criteria, none of which can claim universal validity. It is a matter of design of a procedure according to stated principles, implementation, and ex post examination of statistical properties.

The main approach of ANVUR, as already stated, was to leave the expert panels free to assign different weights to citations and journal indicators in defining the final score for those articles for which there was no correlation between the indicators. As described more in detail in Ancaiani et al. (2015) and Anfossi et al. (2016) and summarized in Bonaccorsi (2020), experts may give more weight to journal-level indicators or to normalized citations in the formula that combines the two metrics in order to assign a class of merit, hence the final score for the item. This choice reflects the overall practice in the scientific community and is justified in the final Report of each GEV.

A serious difficulty that ANVUR had to face was that the Ministerial decree opening the first VQR gave detailed indication of the classes of merit (see Table 1 of the companion paper for details) and of the quantiles associated to each of them. Following this prescription, the square space defined by the percentiles of the two indicators was partitioned into a series of squares whose boundaries were defined by the percentiles of the final classes of merit. These quantiles were not uniform (e.g. quartiles), but variable in their range: in particular, a large region of the distribution (below the median) was to be evaluated as Limited. Under these circumstances, it is easy to demonstrate that small changes in one of the indicators would result in large changes in the final classification. This objection was raised shortly after the first VQR (2004–2010).

The procedure has been the object of criticism (Abramo and D'Angelo, 2015; 2017; Franceschini and Maisano, 2017). Since the initial procedure (2004–2010) forced the assignment of discrete classes to each of the criteria, the variability of the final score was large for borderline cases.

ANVUR addressed this issue and modified the aggregation algorithm. This has been corrected in the 2011–2014 exercise, by making the final score a continuous, not discrete, number. In the second VQR (2011–2014) ANVUR introduced a modification of the aggregation algorithm, illustrated in Anfossi et al. (2016).[4]

My interpretation of the debate is as follows. I find a logical contradiction in some of the criticisms. On the one hand, it is regularly argued that bibliometric evaluation should make use of multiple indicators, and avoid the reliance on single indicators. I find this position well grounded in the literature and methodologically sound (AUBR, 2010; Setti, 2013; Moed, 2007; 2017). On the other hand, however, once one starts to aggregate multiple indicators into a single

of impact in those cases where the paper is too young, field normalizations are not taken into account or if autocitations strongly affect the final evaluation" (Anfossi et al. 2016, 673).

[4] In a nuthshell, the overall square space defined by the range of the percentiles for the two indicators has been divided into oblique strips whose slope is a function of the weights assigned to them by the expert panel. Franceschini and Maisano (2017a; 2017b) criticized this solution, arguing that any non-linear transformation of the two indicators leads to a distortion of the statistical properties of the underlying indicators. Benedetto et al. (2017) replied to this criticism by calculating the proportion of articles that might be misclassified by using the new algorithm (see Benedetto and Setti, 2017 for technical details). They conclude that the effect can be estimated as being 0,05% of the articles that could be submitted by Italian authors, that is, negligible.

evaluation, the argument is raised according to which *any* transformation would create unacceptable distortions, whatever the order of magnitude of the error.

In particular, the arguments of Franceschini and Maisano (2017a), made reference to a paper by Thompson (1993). In this paper it is argued that if percentile ranks are not based on equal scales (which is not the case in our context), then percentiles should never be added or averaged, if one wants to avoid large distortions. In the case of VQR, percentile ranks from citations and from journal indicators are neither added nor simply averaged, but combined in a weighted way, controlling for the resulting error. They are taken as signals of quality, divided in classes of intensity of the signal. The final score is the result of a broad convergence (same class of merit, or intensity of the signal), or the result of a weighted appreciation of the informativeness of the signal (slightly different classes of merit for the two indicators). By modulating the parameters of the aggregation algorithm (see Anfossi et al. 2016 for discussion) it i possible to minimize the errors of classification. It is a matter of measurement, not of admissibility of the procedure. If the signals are contradictory, peer review is adopted.

In using multiple indicators a balance must be defined between sophistication and transparency. The combination of two bibliometric indicators enlarges the basis for judgment. While for individual articles there is some risk of misclassification, the error is very small at aggregate level. The issue of aggregation would be automatically solved if the use of journal-level indicators will be eliminated in the next VQR 2015–2019.

The independent Expert group (2019) suggested to eliminate the classes of merit and use instead five quality criteria, each with 8 points available for evaluation. They also recommended to use one method only within any GEV, to be indicated in the Call. In the case of bibliometrics, they recommend to use only one database, again to be indicated in the Call.

5. Selection of products vs full production of scholars

A few authors argued that it would be better to evaluate all products in a given time window, rather than a sample of self-selected publications (Abramo, D'Angelo and Di Costa, 2014). The criticism was based on two arguments. First, submitting only 3 or 2 items per capita results in a compression of the variability of performance of researchers. Second, researchers may be unable to select their best papers.

The two arguments have grain of truth but do not stand after further reflection. It is certainly true that highly productive researchers publish much more than 3 or 2 papers. However, evaluating all products is not feasible. The first obstacle is obvious: bibliometric evaluation can be done reliably only for STEM disciplines. It would be difficult to defend the idea that scholars in STEM are evaluated on their *full* production, while scholars in SSH are evaluated only for a subsample.

The second obstacle is that for non-indexed items there is no authoritative source of classification that may permit to carry out peer review only on scientific items. The cost of submitting to peer review the entire production of scholars in SSH would be huge. As a matter of fact, the only country that has submitted all products to evaluation is Australia, which decided to evaluate only STEM fields.

As to the second objection, it is true that scholars may make mistakes in their submissions. But, again, several considerations make this issue irrelevant. First, the Italian VQR is by design and legislative mandate an exercise in which *all* researchers must be involved. The early experience of VTR, in which individual researchers had no voice in choosing the products, was not considered positive. With some hindsight, making individual researchers responsible for the selection of products has been an important step for the legitimation of the exercise. Compare this with the recurring controversies in the UK about the filtering role of departments and universities in the decision about who will be subject to evaluation (Sayer, 2015).

Second, researchers may learn about their best papers. Once the main evaluation criteria have been established, researchers learn rapidly which products are to be submitted. As a matter of fact, several researchers and even a university developed software programs (available in Open Source) to advice scholars about their submission choices.

Finally, in all fields in which there are several co-authors, there is a major decision, to be made at university level, in order to avoid double-counting (i.e. the same co-authored paper may be submitted by all affiliations involved, but not twice within the same affiliation). This means that there will be some negotiation between researchers and their Research offices, with the latter adopting an objective function defined in terms of maximization of the overall institutional score.

Summing up, the criticism is not realistic in the context of VQR. It remains true, however, that a major goal for research assessment would be the construction of an official repository of all publications of all researchers. As a matter of fact, ANVUR fully recognized the importance of submitting to assessment the entire scientific production of researchers. The Italian legislation included since 2009 a provision for the full publication of metadata on all products of researchers affiliated to universities, in a National Register of publications.[5] This infrastructure was in the mandate of the Ministry of University and Research (MIUR), since the law delegated the Ministry to develop the implementation. The infrastructure could benefit from the legacy of so called *loginmiur*. As described in the companion paper (Bonaccorsi, 2020), this

[5] The acronym is ANPREPS (Anagrafe nominativa dei professori ordinari e associati e dei ricercatori contenente per ciascun soggetto l'elenco delle pubblicazioni scientifiche prodotte), introduced by Legge 9 gennaio 2009, n. 1.

is a platform managed by a large IT consortium of universities (CINECA), in which all researchers deposit the metadata of all their publications. The platform is routinely used by the Ministry for administrative procedures (e.g. submission of proposals) but it is not open to the public. Immediately after its creation, ANVUR approved a document[6] in which it proposed a framework for the creation of the Register based on loginmiur. It also started a series of technical meetings to accelerate the disclosure of all information in the platform, as the first step towards a validated Register. It turned out that the official obstacle to the publication was the need to define the profile of privacy of researchers, after the recent introduction of a strict legislation. As a matter of fact, this legislative provision is still waiting for implementation, after 11 years.

In the same line of engagement, ANVUR developed a proposal for an experimental facility of collection of metadata in SSH fields, covering non-indexed journals, particularly in national language. A task force[7] developed a technical feasibility proposal, which was shared with the main academic publishers and the largest digital platform for academic journals (Torrossa). The proposal was aimed at establishing a pilot digital platform that would have aggregated and made retrievable the table of contents of a large number of journals in national language and would have extracted automatically the metadata, using machine learning techniques. In this way it would have been possible, after a few years of experimentation, to track the entire journal production of Italian researchers in those fields not covered by bibliometric databases. The proposal included a provision for experimental research on citation extraction from non-indexed sources, for which technical tests had been done with promising results using text mining techniques.

It is useful to remark that this initiative was taken after the feasibility study proposed by Henk Moed (Moed et al. 2009; see also Hicks and Wang, 2009), but much earlier than similar ideas were popularized at European level, particularly by the COST initiative ENRESSH.[8] As a matter of fact, large bibliographic databases in SSH are available mainly in small European countries (Sile et al. 2018), so that the Italian initiative would have been the first one for a large European country. No comparable open repositories are available still today in large European countries. CRIS systems are still not diffused.

The proposal was then discussed in a large conference, held in Rome in 2013 with the editors of academic journals and the scientific societies in all SSH fields. To the surprise of the ANVUR organizers, there was a well organized strong and negative reaction by scientific societies in legal disciplines. The argument was that the proposal was a hidden attempt to introduce bibliometrics into the SSH fields. After collecting metadata, the argument was, ANVUR might have been in the position to start bibliometric analysis against the opinion of scholars. There were even arguments about the legal obstacles to collect data on publications of scholars.

Summing up, the proposal of submitting all publications of scholars to evaluation is valuable in principle, but it must overcome several serious obstacles to become practical. It is a good argument in general, while it cannot be used to dismiss the VQR experience.

6. Cost of the exercise

Another issue that was raised after the start of the VQR was the total cost of the exercise. Geuna and Piolatto (2015) estimated the total cost of the first VQR at approximately 10 million euro (excuding the costs for universities) a level which is in line with the UK experience. In the UK, however, the research assessment is responsible for a larger share of performance-based funding than in Italy. The authors therefore recommended to increase the share of government funding allocated according to merit criteria, in order to justify the expenditure for the assessment procedure.

A careful analysis of the costs of VQR, compared with the REF, has been carried out in the context of the 2018 biannual Report on the State of the research.[9] It is estimated a cost eight times smaller than the REF.

Another estimate, which circulated widely in the media, was proposed by Sirilli (2012).[10] This author assumed that the shadow cost of referee work for the peer review exercise was represented by the average pay at the European Commission level. The resulting estimate is an astonishing 300 million euro total sum. This paper is a nice example of a deeply flawed reasoning. On the one hand, it is hard to believe that the opportunity cost for all researchers, in all fields, for all working days of the year, is equivalent to the daily fee at European Commission. If this were true, than most of academic activities would be immediately halted, since they do not provide this level of income at all.

[6] Delibera n. 5 del 22 giugno 2011. Available at https://www.anvur.it/archivio-documenti-ufficiali/delibera-5-2011-2/.

[7] The composition of the Task force is available at https://www.anvur.it/news/gruppo-di-lavoro-database-e-nuovi-indicatori/.

[8] This important initiative has collected a number of national registers of publications in SSH and suggested technical solutions for their integration in a multilanguage data integration framework. See https://enressh.eu/working-group-3/objectives/ on the COST initiative and https://ecoom.uantwerpen.be/sites/en/edrssh/0/europeandatabasesmap for access to national databases.

[9] Available at https://www.anvur.it/download/rapporto-2018/ANVUR_Rapporto_Biennale_2018_Sezione_10.pdf.

[10] In citing a post from a website I make exception to my methodological principle to comment only published articles in peer-reviewed journals. This is motivated by the large impact of Sirilli's estimate, which was repeated over and over again by critics of the research assessment.

On the other hand, the reasoning violates simple economic principles. All referees of the VQR were asked whether they were prepared to carry out peer review in exchange for a payment of 30 euro. Very few declined. More than 14,000 referees accepted and carried out the work. If all these people considered that their opportunity cost was 450 euro per day, then the vast majority of them would have declined altogether an offer of 30 euro per paper. If most of them accepted it is because they considered the payment adequate. In economic terms, if they were free to accept the offer, their acceptance is a clear indication that they consider the benefits they receive as comparable to their effort. Their revealed preferences, economists would say, mean that the value of their effort is equal to the price received (plus maybe unobserved intrinsic value).

Needless to say, the collaboration of referees was not mainly the result of economic calculations, but most likely the consequence of a feeling of academic obligation and institutional compliance. What does it mean for the estimate of the cost of VQR? According to Sirilli the time of referees is taken away from research and must be evaluated separately as an additional cost. If this argument is true, why are researchers all over the world doing peer review for free? Are they detracting from their research duties? The argument does not stand. Peer review is an integral part of the academic work. Referees learn a lot in doing peer review and have the opportunity to have an impact on the direction of science. The overall cost estimation is therefore grounded on wrong economic assumptions.

7. Journal rating

As stated in the companion paper, in order to produce the thresholds of indicators for the admission of candidates to the National Scientific Habilitation, ANVUR had to classify journals in scientific and non-scientific, and to create a list of A-class journals, to be used for SSH disciplines. Most of these journals were in national language and were non-indexed in the citation indexes Web of Science or Scopus.

As it is well known, journal rating is a controversial issue in the literature. While most authors agree on the notion that scientific journals are different from non scientific ones, the ranking into merit classes is contested. According to some authors, journal rating is a source of conformism that depresses original and unorthodox research (Willmott, 2011; Alvesson and Sandberg, 2013; Mingers and Willmott, 2013; Mingers and Yang, 2017). According to Rafols et al. (2012) journal rating inhibits interdisciplinary research. The Italian experience came after the demise of the journal ranking experiences in Australia and France (Pontille and Torny, 2010). At the same time there were promising experiences in Spain (Giménez-Toledo et al. 2007; 2013) and in several other countries and the expert-based rating of journals was recommended in the literature as a suitable alternative to bibliometrics (Nederhof and Zwaan, 1991; Nederhof, Luwel and Moed, 2001; Hicks and Wang, 2011).

On the practical side, the ranking was anyway mandated by the Ministerial decree with a fixed deadline and there was no time to enter into an extensive consultation.

ANVUR adopted an expert-based classification, in which the opinions of learned societies were a necessary input. The overall approach was reputational, not indicator-based. It was felt that the overall scientific community had a certain agreement on those journals that are essential to the advancement of the disciplines. The expert panel could make use of referees' opinion and use their judgments to support a decision. All learned societies were asked to produce a list of A-rated journals. In the absence of a mandatory upper limit on the total number of journals or percentage of the total in the A class, many learned societies suggested a very large proportion of the total. The short time window made it difficult for some learned societies to contribute. As a matter of fact, some of these opinions were missing while others could not be accepted. The expert panel convened by ANVUR took the responsibility to draft the final decision. Several mistakes were soon discovered.

The publication of journal lists generated a wave of discussion in the academic community. It appeared soon that there was a need to strenghten the dialogue with learned societies and to make the procedure of journal classification open to appeal and renewal. In 2013 a new regulation was issued, in which it was possible for editors of journals to submit new candidatures and to appeal against the rejection into class A, providing new evidence about the reputation in the scientific community. It was stated that the submission of candidatures would be an annual procedure. The process started and several new expert panels, in substitution of the initial one, were nominated. A massive process of examination of new submission started again in 2013 and was kept open at regular intervals.

The issue became even more sensitive in 2016, given that the new procedures for the Habilitation made the threshold much more relaxed, but asked candidates to overcome two out of three thresholds without exceptions. While in the 2012 procedure the members of the committee could choose to depart from the median values by publishing a motivation before the examination of candidates (art. 6, Ministerial Decree n.76/2012), this provision was eliminated in the new procedure. As discussed more in detail in Bonaccorsi (2020), in practice all candidates had a keen interest in the boundaries of the A-class, given that small changes could easily revert the admissibility status. A number of legal actions were taken. This provision forced a process of administrative bureaucratization. An immediate consequence was that the new expert panels, to be nominated in 2017, could not be chosen by ANVUR but were selected after an open call procedure for self-candidature. All steps were subject to the strict requirements of administrative law and with an eye to avoid legal actions.

In evaluating the practice of journal rating an important test is whether the rating of a journal is a good predictor of the quality of the articles that are published in it. This is the classical problem of journal-based indicators, but here we entirely miss the citation-based indicators at article level. An interesting opportunity was offered by the parallel

deployment of VQR and Habilitation. As stated above, the VQR 2004–2010 started in 2011 and was published in 2013. During 2012, therefore, members of the GEV in the non-bibliometric fields were requested to carry out peer review on journal articles submitted by all researchers, irrespective of their academic rank. Exactly in the same period the expert panel was drafting the list of A-class journals for the Habilitation. Within the GEV activities a procedure for journal rating was implemented with the aim to provide external referees with additional information, following the informed peer review approach. A small number of top journals were selected by the GEVs, after consultation with learned societies. These lists were also made available to the expert panels in charge of journal ratings within the Habilitation procedure.

It was then possible to examine to what extent the score assigned to individual articles was correlated with the merit class of the journal. Bonaccorsi et al. (2015) and Ferrara and Bonaccorsi (2016) provided evidence that, controlling for other factors (language of the journal, disciplinary field, academic rank of the author) the probability to receive an Excellent score was twice as large for articles published in A-rated journals as it was for papers in other journals. This argument was criticized by Baccini (2016) according to whom a regression model is not appropriate given that the variables are not independent. Referees who knew about the rating assigned by the GEV were influenced in evaluating individual articles, so that the evaluation at article level is not independent on the evaluation at journal-level. Bonaccorsi et al. (2018) re-examined the issue and offered further arguments. First, the set of A-class journals under the Habilitation was much larger than the small set used for the VQR procedure, three times larger. Second, not all journals rated in A class for the VQR were also rated for the Habilitation. Third, referees under the VQR were instructed to formulate their judgment only after careful reading of the article, using the information on the A-class only as a supporting information. Under the VQR researchers submitted a small selection of their best articles, while for the Habilitation they submitted all their publications (or a large selection). This means that the proportion of articles from top journals was by definition larger for the VQR. Referees had the task of evaluating whether articles published in top journals were indeed excellent, or only good, or even only adequate. Experienced referees know very well that within top journals one can find by definition articles of better *average* quality, but also, more or less occasionally, articles of lower quality or even poor ones. If one were to believe that referees just rate mechanically as excellent all articles published in A-rated journals (whatever the size of the A class) then it would be meaningless to use peer review. Summing up, Bonaccorsi et al. (2018) maintained that between the variables at journal and individual level there was sufficient independence to warrant the regression approach and rejected the argument by Baccini (2016).

Journal rating is still in place. In a publication landscape with more than 15,000 journals in SSH, as witnessed by the initial loginmiur journal list, it was important to intoduce some criteria for quality. When the exercise started, very few journals in SSH had a formal practice of ex ante peer review. After a few years many more journals have adopted the peer review procedure. The competition for entering the A-class motivated many journals to improve editorial policies, boards, and selection criteria. An argument often raised against journal rating is that it might induce comformism and orthodoxy, preventing the birth of new journals or excluding journals outside the mainstream. After several years in line of procedures for the admission of new journals and for the revision of rejection decisions, I can add a skeptical note on this argument. By consulting the lists of journals published in the ANVUR website it seems that the entry of new journals in the last few years has been continuous. More research is clearly needed for a more balanced assessment.

8. Academic promotion

As largely discussed in the companion paper, one of the main effects of the introduction of research assessment has been the utilization of quantitative indicators as admission thresholds for the habilitation of candidates for academic promotion. In that paper I have illustrated the context of the reform, characterized by a long tradition of lack of transparency, in which promotions by seniority, irrespective of the research performance, have traditionally been diffused and the trasparency of promotion criteria has been modest, if any. I refer to the companion paper for references.

The 2010 reform has drastically changed the overall landscape of academic careers, introducing a new system called *Abilitazione Scientifica Nazionale* (ASN), or National Scientific Habilitation, and placing a heavy pressure on the research record. The introduction of thresholds based on the median value of the distribution of indicators was criticized with harsh comments, within a more general argument against metrics. After a few years it is possible to examine whether the introduction of indicators has damaged the academic system and, more generally, what has been the overall impact. In doing so several authors have exploited an unusual level of administrative transparency, insofar as all the documentation of promotion procedures is publicly available (CV and list of publications of candidates; selection criteria adopted by the committee; individual judgments).

I examine here the impact of the new procedure in terms of mitigation of some of the problems that have afflicted the Italian academic promotion system in the past and that have motivated the 2010 reform, that is promotion by: (i) academic connection; (ii) seniority; (iii) non-publication criteria; (iv) gender.

8.1 Promotion by connection

In an ideal world, the existence of personal connections between candidates and the members of evaluation committees should not enter into the promotion decision. In reality, a large literature shows that this is not the case, with connections creating an advantage for some of the candidates.

Exploiting the feature of the Habilitation procedure according to which the composition of evaluation committees and the list of candidates are published on the ASN website, Bagues, Sylos Labini and Zynovieva (2019) have studied the role of connections for the academic promotion. A connection is defined as the presence in the committee of a co-author, a colleague (same university) or a PhD advisor of the candidate. Given the tradition of favoritism largely discussed in the companion paper (which however was also found in other countries such as France and Spain), the expectation of the study was to find a large connection premium, that is, a significantly larger probability of promotion for connected candidates. The conclusions are surprising: "We find that connected candidates are 4.6 p.p. (13%) more likely to qualify. Instead, Zinovyeva and Bagues (2015) find that in the Spanish system of national qualification evaluations, where evaluation reports are not publicized, the (exogenous) presence of a connection in the committee increases candidates' chances of qualifying by around 50%. Similarly, the work of Perotti (2002) suggests that the impact of connections was significantly higher in the evaluation system that was in place previously in Italy" (Bagues, Sylos Labini and Zynovieva, 2019, 96). In other words, the Habilitation system has apparently curbed one of the most lasting attitudes of Italian academia.

8.2 Promotion by seniority

Another distortion that was frequently denounced is the promotion of academic staff who are no longer active in research, only on the basis of seniority.

Marini (2017) examined the results of the Habilitation procedures in physics, engineering, economics and law in order to verify whether the seniority of candidates (i.e. the number of years in the current rank) still plays a large role, as in the past. His conclusions are clear: "Succinctly stated, these (…) findings reveal a system which tends to favor early careers and good publication records regardless of years of service the individual has notched up in his/her current rank" (Marini, 2017, 202). More precisely, in the case of full professors, "despite some disciplinary differences, generally one or two indicators out of three clearly act as the main determinant of the decisions to award or not to award eligibility to apply for a full professorship. Hence, performance, especially in terms of quality and strategic scientific publications, is the key factor in pushing on and climbing the academic career ladder to the top. Furthermore, younger scholars in each position can bypass their older peers, even with the same indicators of productivity" (Marini, 2017, 202).

It seems that the new system has reduced the role of seniority and has placed a premium on the ability of young candidates to produce good research in their early years. Taking into account the long tradition of promotions by seniority, this seems to be an important result.

8.3 Promotion by non-publication criteria

The new system places large emphasis on transparency of promotion criteria based on publications and on the publicity of the CV and the list of publications. This places a severe constraint on the adoption of non-publication criteria, which in the past were used with more discretionary power. Poggi et al. (2019) have examined the entire collection of CVs of candidates to the 2012 ASN procedure (n= 59149 candidates, with 1,910,873 papers), using several machine learning techniques and a dedicated ontology they identify as many as 291 predictors, or attributes of candidates described in CVs that the members of the committee may have used to make the promotion decision. Their model outperforms the state of the art in predicting correctly the final decision. Among the top 15 predictors we find a small set that describes the career of candidates (affiliation, age, maximum number of years with affiliation to the same university, years since the first publication) but the bulk of criteria are all publication-related, describing the articles, the journals and journal categories, and the Impact Factor. Interestingly, the number of citations do not appear among the top criteria. The continuity of publications, on the contrary, as measured by the number of years without any publication, does appear among the top criteria, confirming the attention of committees to being active in research as a condition for promotion.

8.4 Promotion by gender

The notion that academic promotion is affected by gender bias has been examined by a large literature, which I cannot review here. I am interested in understanding whether a more transparent system such as ASN has mitigated the large gender discrimination that has afflicted the Italian academic system (and that of other countries as well), since long time. A few papers have addressed this issue.

De Paola and Scoppa (2015) have shown that promotion committees that include a woman have a higher probability to give promotion to women, controlling for research productivity. Bagues, Sylos Labini and Zinovyeva (2017) find a more complex effect, in which women in the promotion committee give higher scores to female candidates, but having a woman in the committee make the male members more severe against women. According to Marini and Maschitti (2018) the gender discrimination is still large, as men have around 24% more probability to be promoted full professors, in the 2013–2016 period, at parity of scientific production. This effect, however, is not due to the ASN procedure, but to the downstream decentralized process of promotion decided by departments among competing candidates, all having the habilitation. In other word, "evidence tells there is less gender discimination at ASN level, and substantally more gender discimination at the promotion level" (Marini and Maschitti, 2018, 1002). This effect is confirmed by Filandri and Pasqua (2019) who study the probability of career advancement (in both ranks of associate and full professor) in the period 2012–2016 for those professors who were accredited in 2012 or in 2013. They find that "on average,

female assistant professors have a probability of advancement to associate professorships which is 8 percentage points lower than their male colleagues. This difference increases to 17 percentage points when we consider associate professors' probability of becoming full professors" (Filandri and Pasqua, 2019, 12). In other words, while the ASN procedure has reduced the gender discrimination effect, the effect is reproduced at decentralized level, for which the degree of transparency is lower.

Summing up, it seems that the introduction of indicators has reduced the weight of non-academic factors such as seniority and personal connections, has increased the importance of research productivity along the entire career, and has mitigated the gender discrimination at national level. The discrimination did not disappear, however, in the second stage of the procedure at departmental level.

9. Performativity of research evaluation and unintended consequences

9.1 Behavioral impact

There is little doubt that, given the pervasiveness and the impact of research assessment at institutional level, it has influenced the behavior of researchers. This is even more so given the joint introduction of evaluation at university level and of indicators in the recruitment process. A trickle down is certainly in place (Aagard, 2015).

According to Moed (2007) and Mingers and Leydesdorff (2015) in the application of bibliometric indicators for research evaluation it is important to put in the agenda the issue of behavioral impact. After being created, indicators take a life of their own. Users of indicators throw away the instructions for use and select quick-and-dirty information for their convenience (van Raan, 2005). In turn, the very existence of quantitative indicators has large implications for the behavior of researchers. Indicators create incentives and disincentives, recommend some behaviors and discourages others (Burrows, 2012; Dahler Larsen, 2014). A recent literature has called the attention on the unintended consequences (Weingart, 2005) on the behavior of researchers. Examples of negative impact include goal displacement (i.e. aiming at meeting indicators, not producing valuable knowledge), selection of communication channels that are not consistent with the scientific community (e.g. publishing in English journals instead of writing books in national language), or lack of integrity (e.g. engaging into gift or cohercive authorship, gaming with citations, and the like) (Laudel and Gläser, 2006; Van Dalen and Henkens, 2012; Hammarfelt and de Rijcke, 2015; De Rijcke et al. 2016; Muller and De Rijcke, 2017). Following this critical literature a few authors have argued that the introduction of research assessment in Italy has produced negative behavioral consequences.

Baccini, De Nicolao and Petrovich (2019) show that after the introduction of the research assessment the number of citations by Italian authors to other Italian authors increased significantly and much more than in other countries. They interpret this evidence as a perverse effect of evaluation, leading researchers to inflate artificially their citations in order to favour colleagues (and, implicitly, expecting reciprocity). Overall, I find this paper flawed in providing compelling evidence and rigorous counterfactual reasoning. It is not at all demonstrated that the increase in citations to Italian colleagues is due to perverse gaming with indicators. First, according to their data the process started around 2010, i.e. before or shortly after the introduction of research assessment. Thus causality assumptions are weak. Second, making gift citations to other colleagues may be risky if the citing and the cited authors are in competition in promotion procedures. It is not clear why Italian researchers should be generous with their competitors. Third, it is likely that the overall internationalization and quality of Italian researchers increased in the period (perhaps as an effect of research assessment?), so that part of the citations are genuine recognition of merit, not gaming. The paper suggests a strong causality effect, without demonstrating it.

A more controlled counterfactual approach is taken by Seeber et al. (2019) in showing the increase in self-citations in four disciplines as a consequence of the introduction of bibliometric indicators in the National Habilitation. Their results are credible and point to gaming with citations when candidates are at the borders of classes of merit. At the same time, one might argue that the convenience of gaming depends very much on the expected outcome. In particular, it is easier to game with citations than with journal factors, and it is easier to game when the thresholds are low. In other words when the admissibility threshold to become associate professor is, say, an h-index of 4, it is easier to organize a clique of reciprocating citers in a few years and overcome the threshold. If the h-index is, say 10 or 12, the game is much more difficult.[11] Therefore this evidence is less dramatic than it is often said. It points to the need to design incentives in a way to anticipate gaming behavior, not necessarily to avoid indicators at all. As an example it is easy to introduce a control for self-citations and to eliminate them in any consideration of indicators.

More generally, it should be remarked that this literature is largely based on case studies, university-level field observation, and conceptualizations. As De Rijcke and co-authors are led to conclude after their extensive survey, "many studies are of tentative and theoretical nature, prophesizing on potential effects rather than documenting actual

[11] Incidentally, this is a strong reason for me to argue that the abolishment of medians has not been good. The medians were criticized with several arguments (Marzolla, 2015), some of which were well grounded. But the weakening of the thresholds may have unintended consequences as well. For example, gaming around the threshold is relatively easy, while gaming around the *median* value of any distribution of indicators is really difficult, give the robustness of the moment of the distribution.

consequences" (De Rijcke et al, 2016, 6). I share the comment by Sivertsen (2017), who noted that this caution is completely lost in the reception of the *Metric tide* report (Wilsdon et al. 2015) and in the subsequent debate.

Much more research is needed before concluding that the behavioral impact of research assessment is detrimental to the scientific enterprise.

9.2 Deterioration of pluralism

Among the negative consequences of research assessment the reduction of epistemic pluralism is often mentioned (Viola, 2017). Given the role played in research assessment by citation indicators, there is inevitably a premium on research topics that are more popular and journals that have larger impact factors. In turn, this creates a disadvantage for minority positions.

A case in point is represented by economics, in which minority positions are represented by non-mainstream scholars, working in various non-neoclassical traditions (Marxist, Sraffian, Austrian, Post-keynesian, or institutional economics). Corsi, D'Ippoliti and Zacchia (2018; 2019) have repeatedly criticized the research assessment procedures carried out in Italy as a source of discrimination against heterodox economists. In a *Research Policy* paper they examine the habilitation decisions in the field of Economics and find a number of factors that led to negative outcomes. Among them are the number of books and the number of articles or chapters in books (two of the three threshold indicators) and the number of articles in heterodox journals. Promotion committees gave the habilitation almost exclusively on the basis of the number of articles published in A-rated journals, or a list of top 454 journals, irrespective of the fact that heterodox candidates had produced more articles in non-top journals, chapters and books than their mainstream colleagues (Corsi, D'Ippoliti and Zacchia, 2019). They use this evidence to join the opinion, held by several heterodox scholars in the UK after the RAE/REF experience, according to which research assessment is responsible for the elimination of dissenting views.

I believe that research assessment should be neutral with respect to epistemic differences in disciplines (Bonaccorsi, 2018b). In practical terms, scholars from mainstream and minority positions should be systematically involved in expert panels and committees. This is what happened in the first VQR 2004–2010, in which the GEV in Economics was chaired by a leading mainstream economist (Tullio Jappelli) but included, among others, the leaders of heterodox economics in the institutional (Neri Salvadori) and evolutionary (Giovanni Dosi) traditions. Interestingly, while after the 2000–2003 VTR there was a famous minority document from Luigi Pasinetti (a leading authority in structural and Austrian tradition) contesting the evaluation, in the 2004–2010 GEV all judgments and scores were approved with unanimous vote. When addressing the literature that criticizes the research assessment as a threat to pluralism, I find some arguments problematic in the causal assumptions. Scientific disciplines have a dual dynamics, one of epistemic type in which competing theories and paradigms challenge each other to address relevant scientific issues, and one of sociological and institutional type, in which the reproduction of scholarship is at stake. It is not clear whether the relative dynamics between mainstream and heterodox positions depends causally on research assessment. In order to substantiate this argument one should at least show that in countries in which there is no research assessment heterodox positions survive better or grow more than in countries subject to research assessment.

9.3 Polarisation of the higher education system

The publication of the results of VQR 2004–2010 showed a large gap between universities located in the Southern regions and those located in the North and Centre. Southern universities are placed, with limited exceptions, in the bottom part of the ranking used to allocate performance-based funding. Although the financial implications have been mitigated by the Ministry by placing an upper limit on the penalization, the impact has been almost immediate.

This has opened a large debate on the risk that the gap could be widened and made irreversible. Southern universities, which operate in less privileged areas and are subject to large student-staff ratios due to the demographic pressure, the argument goes, will receive less and less resources and will never be in the position to improve their position. Several authors have therefore argued that research assessment in the Italian context is an instrument for perpetuating and widening spatial and social inequalities (Viesti, 2016; Grisorio and Prota, 2020). This issue is part of a more general argument according to which research assessment is invariably associated to the deepening of inequalities (Warren et al. 2020).

This argument deserves close attention. The premise of research assessment is that it can produce behavioral changes that improve the position of those below the average. Allocating resources according to the quality of research should create appropriate incentives for improving, either by placing more effort in research and by recruiting academic staff with a good research record. If, on the contrary, financial constraints or resource endowments make the improvement unlikely for poor performers, the gaps become irreversible. The importance of these dynamic effects must be examined empirically, however. A crucial issue here is that the introduction of research assessment and performance-based funding has taken place in Italy in parallel with significant cuts in the government budget, particularly after the 2008 crisis. This is a major government and political mistake, insofar it has created the deeply held belief that performance-based funding is nothing else than a technical instrument to reduce the resources to the higher education system. Under these conditions, even marginal reductions in funding to universities in Southern regions may be serious.

From this perspective, it can be said that research assessment is held responsible for others' faults. From an empirical perspective, however, the depauperation effect of research assessment is not supported. Following the methodology

introduced by Buckle et al. (2020) in New Zealand, Checchi et al. (2020) found that the polarisation effect is not confirmed by the empirical evidence. Comparing the research quality of universities between the VQR 2004–2010 and the VQR 2011–2014, after making scores comparable, they find a remarkable process of convergence towards the mean, or reduction of inequalities between Southern universities and universities located in North and Centre.

More empirical work must be done, however, to examine the unintended consequences on spatial and social inequalities, particularly under a regime of fiscal discipline.

10. Policy highlights and conclusions

In a relatively short time frame the Italian research and higher education systems have been subject to a pervasive introduction of evaluation. There is probably no other sector of the public administration in which all personnel is subject to evaluation in such a systematic way.

It is no surprise that the introduction of research assessment has generated a huge debate. In this paper I have distilled the most relevant criticism, from a technical and substantive perspective.

I find some of the criticism well founded and constructive. For example, it is clear that the use of journal-based indicators and the need to build up a weighting scheme are problematic from the perspective of the state of the art of evaluative informetrics (Moed, 2017; 2020). Nevertheless, they are used quite frequently in evaluation studies, according to a recent meta-analytic survey (Jappe, 2020).

I believe the methodological choices used in research assessment should be the object of open discussion. Once they are adopted, it is good policy to keep them stable for a certain number of years, in order to align the behaviours of researchers and ensure comparability of the exercises over the years. This implies some stickiness and the ability to justify the choices under the fire of criticisms.

At the same time, things may change. To make an example, if in the future the journal-level indicators will be eliminated, that will not imply any loss of legitimacy of the previous choices. Pragmatically, it would be useful to compare the results with an appropriate sampling approach in order to derive policy implications.

On the contrary, the criticism against the dual methodology (peer review and bibliometrics) and the argument in favour of the assessment of the entire production of scholars, instead of a submitted sample, are not realistic.

Using peer review for SSH (with a few exceptions) and bibliometrics for STEM is a methodological choice that can be defended, after appropriate normalization, in a large scale exercise. Given the impossibility to submit all products to peer review due to budget constraints, the choice to ask researchers to submit a selection of products of their choice is also good practice. Even if the submission is done not by individuals but by universities, the involvement of all researchers creates more engagement.

At the same time, the arguments that research assessment promotes research misconduct, reduces epistemic pluralism and interdisciplinarity, does not address gender discrimination, and deepens regional inequalities are conceptually and methodologically weak.

As stated above, the premise of research assessment is that those below the average may find incentives, opportunities and guidelines to improve their research performance. The overall system should monitor closely whether this happens, or whether dynamic self-reinforcing mechanisms widen inequalities and made them irreversibile. To the best of my knowledge, the argument that research assessment is damaging the scientific system and overall society is not empirically well grounded.

Research assessment is by nature and purpose a social experimentation, ultimately rooted in the scientific method, hence open to criticism. Even harsh criticism. At the same time, there is no perfect research assessment. The argument "better no assessment than imperfect assessment" is ideological. We should pursue improvement, not perfection.

I hope the paper has offered a dispassionate discussion.

Competing Interests

The author declares no competing interests. The author has been member of the Board of ANVUR in the period May 2011–May 2015.

References

Aagard, K. (2015). How incentives trickle down. Local use of a national bibliometric indicator system. *Science and Public Policy*, 1–13.

Abramo, G., D'Angelo, C. A., & **Di Costa, F.** (2010). Citations versus Journal Impact Factor as proxy of quality: Could the latter ever be preferable? *Scientometrics*, *84*, 821–33. DOI: https://doi.org/10.1007/s11192-010-0200-1

Abramo, G., D'Angelo, C. A., & **Di Costa, F.** (2011). National research assessment exercises. A comparison of peer review and bibliometrics rankings. *Scientometrics*, *89*, 929–941. DOI: https://doi.org/10.1007/s11192-011-0459-x

Abramo, G., D'Angelo, C. A., & **Di Costa, F.** (2014) Inefficiency in selecting products for submission to national research assessment exercises. *Scientometrics*, *98*, 2069–2086. DOI: https://doi.org/10.1007/s11192-013-1177-3

Abramo, G., & **D'Angelo, C. A.** (2015). The VQR, Italy's second national research assessment: Methodological failures and ranking distortions. *Journal of the Association for Information Science and Technology*, *66*, 2202–2214. DOI: https://doi.org/10.1002/asi.23323

Abramo, G., & **D'Angelo, C. A.** (2017). On tit for tat: Franceschini and Maisano versus ANVUR regarding the Italian research assessment exercise VQR 2011–2014. *Journal of Informetrics*, *11*(3), 783–787. DOI: https://doi.org/10.1016/j.joi.2017.06.003

Alvesson, M., & **Sandberg, J.** (2013). Has management studies lost its way? ideas for more imaginative and innovative research. *Journal of Management Studies*, *50*(1), 128–152. DOI: https://doi.org/10.1111/j.1467-6486.2012.01070.x

Ancaiani, A., Anfossi, A. F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., Cicero, T., Ciolfi, A., Costa, F., Colizza, G., Costantini, M., di Cristina, F., Ferrara, A., Lacatena, R. M., Malgarini, M., Mazzotta, I., Nappi, C. A., Romagnosi, S., & **Sileoni, S.** (2015). Evaluating scientific research in Italy: The 2004–10 research evaluation exercise. *Research Evaluation*, *24*, 242–255. DOI: https://doi.org/10.1093/reseval/rvv008

Anfossi, A., Ciolfi, A., Costa, F., Parisi, G., & **Benedetto, S.** (2016). Large-scale assessment of research outputs through a weighted combination of bibliometric indicators. *Scientometrics*, *107*, 671–683. DOI: https://doi.org/10.1007/s11192-016-1882-9

Archambault, E., Vignola-Gagne, E., Cote, G., Lariviére, V., & **Gingras, Y.** (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, *68*(3), 329–342. DOI: https://doi.org/10.1007/s11192-006-0115-z

AUBR. (2010). Assessment of University-Based Research Expert Group (AUBR). Assessing Europe's University-Based Research. European Commission, Brussels.

Baccini, A. (2016). Reader comment to Bonaccorsi et al. (2015). *F1000Res*, *4*. 196. DOI: https://doi.org/10.12688/f1000research.6478.1

Baccini, A., & **De Nicolao, G.** (2016). Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. *Scientometrics*, *108*, 1651–1671. DOI: https://doi.org/10.1007/s11192-016-1929-y

Baccini, A., & **De Nicolao, G.** (2017). A letter on Ancaiani et al. 'Evaluating scientific research in Italy: the 2004–10 research evaluation exercise'. *Research Evaluation, 26*, 353–357. DOI: https://doi.org/10.1093/reseval/rvx013

Baccini, A., De Nicolao, G., & **Petrovich, E.** (2019). Citation gaming induced by bibliometric evaluation: A country-level comparative analysis. *PLoS ONE*, *14*, e0221212. DOI: https://doi.org/10.1371/journal.pone.0221212

Bagues, M., Sylos-Labini, M., & **Zinovyeva, N.** (2017). Does the gender composition of scientific committees matter? *American Economic Review*, *107*(4), 1207–1238. DOI: https://doi.org/10.1257/aer.20151211

Bagues, M., Sylos-Labini, M., & **Zinovyeva, N.** (2019). Connections in scientific committees and applicants' self-selection: Evidence from a natural randomized experiment. *Labour Economics*, *58*, 81–97. DOI: https://doi.org/10.1016/j.labeco.2019.04.005

Benedetto, S., Checchi, D., Graziosi, A., & **Malgarini, M.** (2017). Comments on the paper "Critical remarks on the Italian assessment exercise", Journal of Informetrics, 11(2017) and pp. 337–357. *Journal of Informetrics*, *11*, 622–624. DOI: https://doi.org/10.1016/j.joi.2017.03.005

Benedetto, S., & **Setti, G.** (2017). Una Analisi Empirica dell'Algoritmo di Classificazione Bibliometrica della VQR2011–2014. www.lavoce.info

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., & **Peracchi, F.** (2015). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Research Policy*, *44*, 451–466. DOI: https://doi.org/10.1016/j.respol.2014.08.004

Bonaccorsi, A., Cicero, T., Ferrara, A., & **Malgarini, M.** (2015). Journal ratings as predictors of articles quality in Arts, Humanities and Social Sciences: An analysis on the Italian Research Evaluation Exercise. *F1000Res*, *4*(196). DOI: https://doi.org/10.12688/f1000research.6478.1

Bonaccorsi, A., Daraio, C., Fantoni, S., Folli, V., Leonetti, M., & **Ruocco, G.** (2017). Do Social Sciences and Humanities behave like life and hard sciences? *Scientometrics*, *112*, 607–653. DOI: https://doi.org/10.1007/s11192-017-2384-0

Bonaccorsi, A. (Ed.) (2018a). *The evaluation of research in Social Sciences and Humanities. Lessons from the Italian experience*. Cham, Switzerland: Springer. DOI: https://doi.org/10.1007/978-3-319-68554-0

Bonaccorsi, A. (2018b). Towards an epistemic approach to evaluation in SSH. In A. Bonaccorsi (Ed.), *The evaluation of research in Social Sciences and Humanities. Lessons from the Italian experience* (pp. 253–267). Cham, Switzerland: Springer. DOI: https://doi.org/10.1007/978-3-319-68554-0_1

Bonaccorsi, A. (2020). Two Decades of Experience in Research Assessment in Italy. *Scholarly Assessment Reports*, *2*(1): 16. DOI: https://doi.org/10.29024/sar.27

Bonaccorsi, A., Ferrara, A., & **Malgarini, M.** (2018). Journal ratings as predictors of article quality in Arts, Humanities, and Social Sciences: An analysis based on the Italian research evaluation exercise. In A. Bonaccorsi (Ed.), *The evaluation of research in Social Sciences and Humanities. Lessons from the Italian experience* (pp. 253–267). Cham, Switzerland: Springer. DOI: https://doi.org/10.1007/978-3-319-68554-0_11

Buckle, R. A., Creedy, J., & **Gellell, N.** (2020). Is external research assessment associated with convergence or divergence of research quality across universities and disciplines? Evidence from the PBRF process in New Zealand. *Applied Economics*, *52*(36), 3919–3952. DOI: https://doi.org/10.1080/00036846.2020.1725235

Burrows, R. (2012). Living with the h-index? Metric assemblages in the contemporary academy. *The Sociological Review*, *60*(2), 355–372. DOI: https://doi.org/10.1111/j.1467-954X.2012.02077.x

Checchi, D., Mazzotta, I., Momigliano, S., & **Olivanti, F.** (2020). Convergence or polarisation? The impact of research assessment exercies in the Italian case. *Scientometrics*, *124*, 1439–1455. DOI: https://doi.org/10.1007/s11192-020-03517-2

Corsi, M., D'Ippoliti, C., & **Zacchia, G.** (2018). A case study of pluralism in Economics. The heterodox glass ceiling in Italy. *Review of Political Economy*, *30*(2), 172–189. DOI: https://doi.org/10.1080/09538259.2018.1423974

Corsi, M., D'Ippoliti, C., & **Zacchia, G.** (2019). Diversity of background and ideas. The case of research evaluation in economics. *Research Policy*, *48*, 103820. DOI: https://doi.org/10.1016/j.respol.2019.103820

Dahler-Larsen, P. (2014). Constitutive effects of performance indicator systems. *Public Management Review*, *16*(7), 969–986. DOI: https://doi.org/10.1080/14719037.2013.770058

De Paola, M., & **Scoppa, V.** (2015). Gender discrimination and evaluators' gender: Evidence from Italian academia. *Economica*, *82*, 162–188. DOI: https://doi.org/10.1111/ecca.12107

De Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & **Hammarfelt, B.** (2016). Evaluation practices and effects of indicator user. A literature review. *Research Evaluation*, *25*, 161–169. DOI: https://doi.org/10.1093/reseval/rvv038

DORA. (2012). *San Fransisco Declaration on Research Assessment*. Available at http://www.ascb.org/dora/.

Fassari, L. G., & **Valentini, E.** (Eds.) (2020). *I sociologi e la valutazione dell'università*. Roma: Carocci.

Ferrara, A., & **Bonaccorsi, A.** (2016). How robust is journal rating in Humanities and Social Sciences? Evidence from a large-scale, multi-method exercise. *Research Evaluation*, *25*(3), 279–291. DOI: https://doi.org/10.1093/reseval/rvv048

Filandri, M., & **Pasqua, S.** (2019). 'Being good isn't enough': Gender discrimination in Italian academia. *Studies in Higher Education*. DOI: https://doi.org/10.1080/03075079.2019.1693990

Franceschini, F., & **Maisano, D.** (2017a). Critical remarks on the Italian research assessment exercise VQR 2011–2014. *Journal of Informetrics*, *11*(2), 337–357. DOI: https://doi.org/10.1016/j.joi.2017.02.005

Franceschini, F., & **Maisano, D.** (2017b). A rejonder to the comments of Benedetto et al. on the paper "Critical remarks on the Italian research assessment exercise VQR 2011–2014" (Journal of Informetrics 11(2), 337–357). *Journal of Informetrics*, *11*, 645–646. DOI: https://doi.org/10.1016/j.joi.2017.05.013

Geuna, A., & **Piolatto, M.** (2015). Research assessment in the UK and Italy: Costly and difficult, but probably worth it (at least for a while). *Research Policy*, *45*, 260–271. DOI: https://doi.org/10.1016/j.respol.2015.09.004

Giménez-Toledo, E. (2020). Why books are important in the scholarly communication system in social sciences and humanities. *Scholarly Assessment Reports*, *2*(1), 6. DOI: https://doi.org/10.29024/sar.14

Giménez-Toledo, E., Román-Román, A., & **Alcain-Partearroyo, D.** (2007). From experimentation to coordination in the evaluation of Spanish scientific journals in the humanities and social sciences. *Research Evaluation*, *16*(2), 137–148. DOI: https://doi.org/10.3152/095820207X220409

Giménez-Toledo, E., Mañana-Rodríguez, J., & **Delgado-López-Cózar, E.** (2013). Quality indicators for scientific journals based on expert opinion. http://arxiv.org/ftp/arxiv/papers/1307/1307.1271.pdf

Glänzel, W., & **Moed, H. F.** (2002). Journal impact measures in bibliometric research. *Scientometrics*, *53*, 171–193. DOI: https://doi.org/10.1023/A:1014848323806

Grisorio, M. J., & **Prota, F.** (2020). Italy's national research assessment. Some unpleasant effects. *Studies in Higher Education*, *45*(4), 736–754. DOI: https://doi.org/10.1080/03075079.2019.1693989

Group of experts. (2019). Report of the Group of experts charged by ANVUR to advice on the process "Valutazione della qualità della ricerca (VQR)". *An independent assessment of the past VQRs carried out by ANVUR.* Rome, 12 March 2019.

Hammarfelt, B., & **de Rijcke, S.** (2015). Accountability in context. Effects of research evaluation systems on publication practices, disciplinary norms and individual working routines in the Faculty of Arts at Uppsala University. *Research Evaluation*, *24*(1), 63–77. DOI: https://doi.org/10.1093/reseval/rvu029

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & **Rafols, I.** (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*, 429–431. DOI: https://doi.org/10.1038/520429a

Hicks, D. (2004). The four literatures of social science. In H. F. Moed, W. Glänzel & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 473–496). Dordrecht: Kluwer Academic Press. DOI: https://doi.org/10.1007/1-4020-2755-9_22

Hicks, D., & **Wang, J.** (2009). Towards a bibliometric database for the Social Sciences and Humanities. A European Scoping Project. *Final Report on Project for the European Science Foundation.*

Hicks, D., & **Wang, J.** (2011). Coverage and overlap of the new social science and humanities journal lists. *Journal of the American Society for Information Science and Technology*, *62*(2), (2011): 284–294. DOI: https://doi.org/10.1002/asi.21458

Hunt, G. E., Cleary, M., & **Walter, G.** (2010). Psychiatry and the Hirsch h-index: The relationship between Journal Impact Factors and accrued citations. *Harvard Review of Psychiatry*, *18*(4), 207–19. DOI: https://doi.org/10.3109/10673229.2010.493742

Institute of Electric and Electronic Engineers. (2013). Appropriate Use of Bibliometric Indicators for the Assessment of Journals, Research Proposals, and Individuals. *Adopted by the IEEE Board of Directors 9 September 2013*.

Jappe, A. (2020). Professional standards in bibliometric research evaluation? A meta-evaluation of European assessment practice 2005–2019. *PLoS ONE, 15*(4): e0231735. DOI: https://doi.org/10.1371/journal.pone.0231735

Jarwal, S. D., Brion, A. M. & **King, M. L.** (2009). Measuring research quality using the journal impact factor, citations and 'Ranked Journals': blunt instruments or inspired metrics? *Journal of Higher Education Policy and Management*, *31*(4), 289–300. DOI: https://doi.org/10.1080/13600800903191930

Larivière, V., Archambault, E., Gingras, Y., & **Vignola-Gagné, E.** (2006). The place of serials in referencing practices. Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science*, *57*(8), 997–1004. DOI: https://doi.org/10.1002/asi.20349

Larivière, V., & **Sugimoto, C.** (2019). The Journal Impact Factor: A brief history, critique, and discussion of adverse effects. In W. Glanzel, H. F. Moed, U. Schmoch & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 3–23). Cham, Switzerland: Springer Nature. DOI: https://doi.org/10.1007/978-3-030-02511-3_1

Laudel, G., & **Gläser, J.** (2006). Tensions between evaluations and communication practices. *Journal of Higher Education Policy and Management*, *28*(3), 289–295. DOI: https://doi.org/10.1080/13600800600980130

Leydesdorff, L. (2008). Caveats for the use of citation indicators in research and journal evaluation. *Journal of the American Association for Information Science and Technology*, *59*(2), 278–287. DOI: https://doi.org/10.1002/asi.20743

Lozano, G. A., Lariviére, V., & **Gingras, Y.** (2012). The weakening relationship between the impact factor and papers' citations in the digital age. *Journal of the American Society for Information Science*. DOI: https://doi.org/10.1002/asi.22731

Marini, G. (2017). New promotion patterns in Italian universities: Less seniority and more productivity? Data from ASN. *Higher Education*, *73*, 189–205. DOI: https://doi.org/10.1007/s10734-016-0008-x

Marini, G., & **Maschitti, V.** (2018). The trench warfare of gender discrimination: Evidence from academic promotions to full professor in Italy. *Scientometrics*, *115*, 989–1006. DOI: https://doi.org/10.1007/s11192-018-2696-8

Marx, W., & **Bornmann, L.** (2013). Journal impact factor: "the poor mans' citation analysis" and alternative approaches. *European Science Editing*, *39*(3), 62–63.

Marzolla, M. (2015). Quantitative analysis of the Italian National Scientific Qualification. *Journal of Informetrics*, *9*, 285–316. DOI: https://doi.org/10.1016/j.joi.2015.02.006

Mingers, J., & **Leydesdorff, L.** (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, *246*, 1–19. DOI: https://doi.org/10.1016/j.ejor.2015.04.002

Mingers, J., & **Willmott, H.** (2013). Taylorizing business school research. on the "one best way" performative effects of journal ranking lists. *Human Relations*, *66*(8), 1051–1073. DOI: https://doi.org/10.1177/0018726712467048

Mingers, J., & **Yang, L.** (2017). Evaluating journal quality. A review of journal citation indicators and ranking in business and management. *European Journal of Operational Research*, *257*(1), 323–337. DOI: https://doi.org/10.1016/j.ejor.2016.07.058

Moed, H. F., Linmans, J., Nederhof, A., Zuccala, A., Illescas, C. L., & **de Moya Anegon, F.** (2009). Options for a Comprehensive Database of Research Outputs in Social Sciences and Humanities. *Standing Committees for the Social Sciences and the Humanities of the European Science Foundation (ESF).* Available at: http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/annex_2_en.pdf

Moed, H. F. (2005). *Citation Analysis in Research Evaluation*. Dordrecht, Springer.

Moed, H. F. (2017). *Applied Evaluative Informetrics*. Cham, Switzerland: Springer. DOI: https://doi.org/10.1007/978-3-319-60522-7

Moed, H. F., & **van Leeuwen, T. N.** (1996). Impact factors can mislead. *Nature*, *381*(6579), 186–186. DOI: https://doi.org/10.1038/381186a0

Moed, H. F., Luwel, M., & **Nederhof, A. J.** (2002). Towards research performance in the Humanities. *Library Trends*, *50*, 498–520.

Moed. H. F. (2007). The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science and Public Policy*, *34*(8), 575–583. DOI: https://doi.org/10.3152/030234207X255179

Moed, H. F. (2020). Appropriate use of metrics in research assessment of autonomous academic institutions. *Scholarly Assessment Reports*, *2*(1): 1. DOI: https://doi.org/10.29024/sar.8

Muller, R., & **De Rijcke, S.** (2017) Thinking with indicators. Exploring the epistemic impacts of academic performance indicators in the life sciences. *Research Evaluation*, *26*, 157–168. DOI: https://doi.org/10.1093/reseval/rvx023

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review. *Scientometrics*, *66*(1), 81–100. DOI: https://doi.org/10.1007/s11192-006-0007-2

Nederhof, A. J., & **Zwaan, R. A.** (1991). Quality judgments of journals as indicators of research performance in the Humanities and the Social and Behavioral Sciences. *Journal of the American Society for Information Science*, *42*(5), 332–340. DOI: https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<332::AID-ASI3>3.0.CO;2-8

Nederhof, A. J., Luwel, M., & **Moed, H. F.** (2001). Assessing the quality of scholarly journals in Linguistics: An alternative to citation-based journal impact factors. *Scientometrics*, *51*, 241–65. DOI: https://doi.org/10.1023/A:1010533232688

Ochsner, M., Hug, S. E., & **Daniel, H.-D.** (2016). *Research assessment in the Humanities. Towards criteria and procedures*. Dordrecht, Springer. DOI: https://doi.org/10.1007/978-3-319-29016-4

Ochsner, M., Kancewicz-Hoffman, N., Ma, L., Holm, J., Gedutis, A., Šima, K. et al. (2020) ENRESSH Policy Brief Research Evaluation. Figshare. *Online resource*. DOI: https://doi.org/10.6084/m9.figshare.12049314.v1

Perotti, R. (2002). The italian university system: rules vs. incentives. In: *Paper Presented at the First Conference on Monitoring Italy*. ISAE, Rome.

Poggi, F., Ciancarini, P., Gangemi, A., Nuzzolese, A. G., Peroni, S., & **Presutti, V.** (2019). Predicting the results of evaluation procedures of academics. *Peer J Computer Science*, June 21. DOI: https://doi.org/10.7717/peerj-cs.199

Poggi, G. (2014). l confronto basato sul Dipartimento Virtuale Associato e sul "Voto Standardizzato (*The comparison based on Virtual Associated Department and the Standardized Score*). Available at https://www.anvur.it/wp content/uploads/2014/02/Dipartimento%20virtuale%20associato%20e%20voto%20standardizzato%20FINALE.pdf

Poggi, G., & **Nappi, C. A.** (2014). Il Voto standardizzato per l'esercizio VQR 2004–2010. *RIV Rassegna Italiana di Valutazione*, *59*, 34–58. DOI: https://doi.org/10.3280/RIV2014-059003

Pontille, D., & **Torny, D.** (2010). The controversial policies of journal ratings. Evaluating social sciences and humanities. *Research Evaluation*, *19*(5), 347–360. DOI: https://doi.org/10.3152/095820210X12809191250889

Radner, S. (1998). How popular is your paper? An empirical study of the citation distribution. arXiv:cond-mat/9804163v2. DOI: https://doi.org/10.1007/s100510050359

Rafols, I., Leydersdorff, L., O'Hare, A., Nightingale, P., & **Stirling, A.** (2012) How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business and management. *Research Policy*, *41*,1262–82. DOI: https://doi.org/10.1016/j.respol.2012.03.015

Sayer, D. (2015). *Rank hypocrisies: The insult of the REF*. Los Angeles: Sage. DOI: https://doi.org/10.4135/9781473910270

Seeber, M., Cattaneo, M., Meoli, M., & **Malighetti, P.** (2019). Self-citations as strategic response to the use of metrics for career decisions. *Research Policy*, *48*, 478–491. DOI: https://doi.org/10.1016/j.respol.2017.12.004

Seglen, P. O. (1997). Why the Impact Factor of journals should not be used for evaluating research. *British Medical Journal*, *314*(7079), 497–498. DOI: https://doi.org/10.1136/bmj.314.7079.497

Setti, G. (2013). Bibliometric Indicators: Why Do We Need More Than One? *IEEE Access*, *1*, 232–246. DOI: https://doi.org/10.1109/ACCESS.2013.2261115

Sile, L., Pölönen, J., Sivertsen, G., Guns, R., Engels, T. C. E., Arefiev, P., Duškova, M., Faurbaek, L., Holl, A., Kulczycki, E., Macan, B., Nelhans, G., Petr, M., Pisk, M., Soós, S., Stojanovski, J., Stone, A., Sušol, J., & **Teitelbaum, R.** (2018) Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey. *Research Evaluation*, *27*(4), 310–322. DOI: https://doi.org/10.1093/reseval/rvy016

Sirilli, G. (2012). Si può stimare che la VQR costerà 300 Milioni di Euro. E a pagarli sarà l'università. Available at www.roars.it; http://www.roars.it/online/si-puo-stimare-che-la-vqr-costera-300-milioni-di-euro-e-a-pagarli-sara-luniversita/.

Sivertsen, G. (2017). Unique, but still best practice? The Research Excellence Framework (REF) from an international perspective. *Palgrave Communications*, *3*, 17078. DOI: https://doi.org/10.1057/palcomms.2017.78

Thompson, B. (1993). GRE percentile ranks cannot be added or averaged. A position paper exploring the scaling characteristics of percentile ranks, and the ethical and legal culpabilities created by adding percentile ranks in making "high stake" admission decisions. *Paper presented at the Annual Meeting of the Mid-South Educational Research Association*, New Orleans, LA, November 12, 1993.

Todorov, R., & **Glänzel, W.** (1988). Journal citation measures. A concise review. *Journal of Information Science*, *14*, 47–56. DOI: https://doi.org/10.1177/016555158801400106

Traag, V. A. (2019). Inferring the causal effect of journals on citations. arXiv:1912.08648.

Traag, V. A. & **Waltman, L.** (2019). Systematic analysis of agreement between metrics and peer review in the UK REF. *Palgrave Communications*, *5*, 29. DOI: https://doi.org/10.1057/s41599-019-0233-x

Traag, V. A., Malgarini, M., & **Sarlo, S.** (2020). Metrics and peer review agreement at the institutional level. arXiv: 2006.14830v1.

Van Dalen, H. P., & **Henkens, K.** (2012). Intended and unintended consequences of a publish-or-perish culture. A worldwide survey. *Journal of the American Association for Information Science and Technology*, *63*(7), 1282–1293. DOI: https://doi.org/10.1002/asi.22636

Van Leeuwen, T. N., & **Moed, H. F.** (2005). Characteristics of journal impact factors. The effects of uncitedness and citation distribution on the understanding of journal impact factors. *Scientometrics*, *63*(2), 357–371. DOI: https://doi.org/10.1007/s11192-005-0217-z

Van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, *62*(1), 133–143. DOI: https://doi.org/10.1007/s11192-005-0008-6

Viesti, G. (2016). *Università in declino. Un'indagine sugli atenei da Nord a Sud*. Roma: Donzelli Editore.

Viola, M. (2017). Evaluation of research(ers) and its threat to epistemic pluralisms. *European Journal of Analytical Philosophy*, *13*(2), 55–78. DOI: https://doi.org/10.31820/ejap.13.2.4

Waltman, L., & **Traag, V. A.** (2019). Use of the journal impact factor for assessing individual articles need not be statistically wrong. arXiv:1703.02334. DOI: https://doi.org/10.12688/f1000research.23418.1

Warren. S., Starnawski, M., Tsatsaroni, A, Vogopoulou, A., & **Zgaga, P.** (2020). How does research performativity and selectivity impact on the non-core regions of Europe? The case for a new research agenda. *Higher Education*, online 5 June 2020. DOI: https://doi.org/10.1007/s10734-020-00559-6

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, *62*(1), 117–131. DOI: https://doi.org/10.1007/s11192-005-0007-7

Willmott, H. (2011). Journal list fetishism and the perversion of scholarship. Reactivity and the ABS list. *Organisation*, *18*(4), 421–442. DOI: https://doi.org/10.1177/1350508411403532

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & **Johnson, B.** (2015). *Metric Tide: Report of the Independent Review of the role of metrics in research assessment and management*. London: Higher Education Funding Council for England. DOI: https://doi.org/10.4135/9781473978782

Zinovyeva, N., & **Bagues, M.** (2015). The role of connections in academic promotions. *American Economic Journal*, *7*(2), 264–292. DOI: https://doi.org/10.1257/app.20120337

How to cite this article: Bonaccorsi, A. (2020). Two Decades of Research Assessment in Italy. Addressing the Criticisms. *Scholarly Assessment Reports*, 2(1): 17. DOI: https://doi.org/10.29024/sar.28

Submitted: 15 September 2020 **Accepted:** 10 November 2020 **Published:** 23 November 2020

Copyright: © 2020 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

Levy Library Press *Scholarly Assessment Reports* is a peer-reviewed open access journal published by Levy Library Press. **OPEN ACCESS**