Journal Ratings as Predictors of Article Quality in Arts, Humanities, and Social Sciences: An Analysis Based on the Italian Research Evaluation Exercise

Andrea Bonaccorsi, Antonio Ferrara, and Marco Malgarini

1 Introduction

There is widespread agreement that research assessment in Social sciences and Humanities (SSH) is complex for a variety of reasons. There are fundamental differences with respect to Science, Technology, Engineering and Mathematics (STEM) fields, especially in the academic publication's structure. In SSH books, book chapters and monographs represent a significant, sometimes dominant, share of scientific production, while journal articles are less central, and national languages are widely used (Finkenstaedt 1990; Hammarfelt 2012). The average number of publications per author, and of authors per publication, is smaller and the time window of citations much longer, so that citations are less immediately useful as indicators of the quality or impact of a publication and the bibliometric approach is thus of limited usefulness (Nederhof et al. 1989). A further complication stems from the number of research quality criteria which is larger in SSH than in other fields. There is less widespread agreement on these criteria (Hemlin 1996; Hemlin and Gustafsson 1996; Ochsner et al. 2012, 2013; Hug et al. 2013, 2014).

This chapter is a revised and expanded version of Bonaccorsi A., Cicero T., Ferrara A. and Malgarini M., *Journal ratings as predictors of articles quality in Arts, Humanities, and Social Sciences: an analysis based on the Italian Research Evaluation Exercise* [version 1; referees: 3 approved]. F1000Research 2015, 4:196 (doi: 10.12688/f1000research.6478.1).

A. Bonaccorsi (✉)
DESTEC, University of Pisa, Pisa, Italy

IRVAPP-FBK, Trento, Italy
e-mail: a.bonaccorsi@gmail.com

A. Ferrara • M. Malgarini
Agenzia Nazionale Valutazione del sistema Universitario e della Ricerca (ANVUR), Rome, Italy
e-mail: antonio.ferrara@anvur.it; marco.malgarini@anvur.it

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_11

The state-of-the-art for SSH research assessment at an international level shows how several roads have been taken to face these challenges. There is widespread agreement that peer review remains the most critical evaluation methodology, and significant efforts are being made to render it more sophisticated, methodologically controlled, based on sound principles of evaluation methodology in social sciences, and free from unwanted biases, distortions, and unexpected side effects. Under this agenda, issues such as the notion of originality, unorthodox science, or interdisciplinarity are under examination (Guetzkow et al. 2004; Hammarfelt 2011).

Much effort is devoted to the classification and evaluation of non-indexed journals (in national languages), which are one of the primary vehicles for academic communication. However, the existing bibliometric databases are limited. The SSH case, which suffers from a scant coverage of relevant publications, partly due to a limited overlap between citing and cited documents (Frost 1979; Hammarfelt 2016), has been carefully discussed in the literature (Nederhof and Zwaan 1991; Nederhof and Noyons 1992; Archambault et al. 2006; Nederhof 2006; Hellqvist 2010). Whatever the specific metrics and the database adopted, the use of citations as the basis for SSH journal ranking has been subject to severe criticism (Christenson and Sigelman 1985; Campbell et al. 2006; Jarwal et al. 2009).

All of this should not lead to the conclusion that scientific production is divided into two irreducibly separated areas, one of which is, by constitution, subject only to qualitative judgement. The issue of the applicability of quantitative methods to the evaluation of research in SSH is open to debate. Classification of journals has been used in several countries for research assessment, sometimes sparking a heated debate.

2 The Debate on the Classification of Academic Journals

As of now, a relatively large body of literature exists on the issue of journal classification. Elsewhere (Ferrara and Bonaccorsi 2016) we discussed its appropriateness and the limits of citation-based and expert opinion-based journal rankings. Hereafter we will focus on criticisms concerning the impact of journal rankings on national languages, academic publication patterns, interdisciplinarity, paradigmatic pluralism and academic freedom. According to such critiques, the production of indicators might orient researchers towards opportunistic publication behaviour (Butler 2003a, b). The entry of new journals might be made too difficult and expensive (Lamp 2009); the rating of journals might lead to an under-investment in interdisciplinary fields (Rafols et al. 2012) and discourage from the undertaking of risky or unorthodox research. Journal classification may be considered the privilege of mainstream science (Rodriguez-Navarro 2009), so that some of the most interesting articles may not be published in top-rated journals (Starbuck 2005).

While these arguments have elements of truth, they do not intrinsically depend on the construction of indicators, as the latter only make the underlying dynamics of recognition more visible. The tension between normal and revolutionary science does not depend in itself on the academic journal system. There is always a trade-off

between pursuing mainstream science and taking a risk by looking for radically new discoveries (Carayol and Dalle 2007). While for individual authors and in the short term the argument might be valid, in the long run, and for large aggregates, the argument is untenable.

The concern that journal rankings might push researchers to switch to the English language, reducing the expressive role of national languages, which are important to all Humanities, is not supported by the literature. Bolton and Kuteeva (2012) carried out an extensive survey at Stockholm University about the use of English in research and teaching and concluded that while English has become standard in STEM fields, in SSH it was auxiliary to the national language.

Another concern is that the classification of journals might induce scholars to switch their publication habits from monographs to journal articles, often opportunistically, and contribute to the so-called "death of the monograph." This is an urgent and worrisome issue in fields like geography (Ward 2009) and literature (Thompson 2002) as monographs are at the core of research (Glanzel and Schoepflin 1999; Thompson 2002; Williams et al. 2009; Gimenez-Toledo and Roman-Roman 2009). Hammarfelt and De Rijcke (2015) have examined the impact of the evaluation system adopted by the Faculty of Arts at Uppsala University in Sweden. They found a significant increase in the share of peer-reviewed publications, without any distortion in the pattern of humanities publications or any "death of monographs." In their view, reaction to journal rating is local, context-dependent and mediated by disciplinary practices.

The interdisciplinarity impact is difficult to evaluate. Rafols et al. (2012) found that journal ranking in social sciences was detrimental to interdisciplinary research in economic areas. This is indeed a serious problem, worthy of consideration from any agency or body in charge of research assessment.

As for conformism and academic freedom, there is literature arguing that rankings service orthodoxy and the mainstream. This argument is mostly based on anecdotal evidence and disciplinary case studies (e.g. in pedagogical research, see the special issue of *Power and Education* on journal rankings). This observation has magnified the problem that in many cases departments and universities base some of their decisions on indicators of journal rankings, instead of assessments of individual articles (e.g. Hasselback et al. 2000; Lee 2006). For example, Gomez-Mejia and Balkin (1992) found that the number of a faculty member's publications, especially those in top-tier management journals, was a primary determinant of faculty pay (see also Park and Gordon 1996). Van Fleet et al. (2000) examined the use of journal lists by departments and warned against the risk that narrow and idiosyncratic lists might distort faculty research attitudes, particularly those of junior members, leading to detrimental conformism. Macdonald and Kam (2008) developed a radical critique of journal rating, pointing to their circularity and arbitrariness, and arguing that its introduction is a threat to non-orthodox and critical research traditions. Hogler and Gross (2009) advanced this direction and argued that journal rating, and indirectly the rating of business schools, is a device for the ideological manipulation of management education, a process that might be examined in the light of the Marxian theory of commodification.

However, some studies have assessed the overall impact of journal rating on academic communities at a national level. Most of these assessments come from Scandinavian countries where journal rating has been extensively introduced, with an impact on university funding. The so-called "Norwegian model," described by Schneider (2009) and Sivertsen (2016), has been especially influential. Developed between 2003 and 2004, it combined a point-based assessment with a performance-based funding system. It was used for funding allocation for the first time in 2006. The system is based on a two-level classification of all publications ("channels"), the collection of data on the whole scientific production, and the assignment of scores to different classes of products. It was also adopted in Denmark, Finland, and some Swedish universities and it triggered an increase in research both in output and the share of output channelled through level 2, or more prestigious, publications (Aagaard et al. 2015). The system's legitimacy is based on the involvement of scholars in the classification process. Challenges are identified in the areas of transparency, representation in individual committees, and the placement of channels among committees (Ahlgren et al. 2012). Ingwersen and Larsen (2014) carried out a detailed examination of the point-based system's impact in Denmark. This system was introduced after Norway's initial experience finding that research article publication was positively affected by the introduction of performance indicators. The latter has, so far, not resulted in an increase in duplicate or redundant publications or a decline in citation impact. Overall, these studies do not lend support to the pessimistic view that journal rating would damage the independence, freedom and established traditions in the practice of research and style of SSH publishing.

3 The Italian Experience

This paper reports on a large experiment in the classification of journals in SSH carried out in Italy between 2012 and 2014 for the National Scientific Habilitation (*Abilitazione Scientifica Nazionale*: ASN). The exercise was based upon a mandatory legal provision to rate *all* journals, to calculate the overall candidates' academic production as part of the national procedure to become an associate professor or full professor. This exercise asked the National Agency for the Evaluation of Universities and Research Institutes (ANVUR) to evaluate all journals in which *at least one* Italian scholar published *at least one* paper in 2002–2012. This figure was more than 60,000 titles.

While the rating of journals was followed in several national contexts, it is only in the Italian exercise that there is the opportunity to carry out a controlled experiment to test the robustness of journal classification. In fact, two independent evaluations were carried out on the same set of journals. Firstly, a panel of experts classified all journals as academic and non-academic (i.e. popular, professional, technical, cultural and political). Then a subset of academic journals was rated as 'A-class.' This exercise was based on the reputation, esteem, diffusion and impact

of journals – a qualitative, expert-based, reputational basis. Elsewhere we have detailed the entire process (Ferrara and Bonaccorsi 2016).

Individual articles published in those journals were rated by a large number of individual referees as part of a nation-wide research assessment exercise (*Valutazione della Qualità della Ricerca*, VQR 2004–10). At least for the considered disciplines, this exercise relied entirely on peer review, carried out by a large number of referees *external* to the panels in charge of the assessment process (Ancaiani et al. 2015). Experts were expected to grade each article on its own merits – which they did, sometimes assigning low scores to articles published in prestigious journals. While referees could have known about the ratings mentioned above (published while the VQR was still ongoing), there is no evidence that any article received higher scores only because it appeared in an outstanding journal. In those fields where journals were indexed and ranked (not simply rated), expert judgements were stricter than expected and some top class papers were assigned a lower score than that allocated to the same class by the journal ranking (Bertocchi et al. 2015). In addition, the Report from the expert panel in Law explicitly clarified that the indication of rating of journals had no systematic influence on the assessment of individual articles, as discussed by Peruginelli and Faro in this volume. This was further evidence that referees did not succumb to any 'halo effect' emanating from the standing of the journal in which the articles were published.

This does not negate that the quality of the peer review process can sometimes be an issue. This problem is not entirely new but has taken a different order of magnitude recently, due to the acceleration of scientific production and the peer-review burden placed upon researchers. Several studies on the peer review process have drawn attention to a variety of biases (Chubin and Hackett 1990; Cicchetti 1991; Daniel 1993; Campanario 1998a, b). Studies by Mahoney (1977) and Travis and Collins (1991) suggested that the familiarity of referees with a subject domain and the affiliation to schools of thought might affect their decisions due to cognitive and institutional particularism. The latter investigated the extent to which experts were biased by their cognitive cronyism, or allegiance to a specific view of scientific practice, irrespective of how professionally they carried out the peer review. This issue has been extensively discussed, but the empirically supported conclusion is that particularism does not dominate in scientific judgement (Cole et al. 1981; Cole 1992). More recently there is the issue of distortions in the peer-review process induced by the need to control costs. Gläser and Laudel (2005) identified several problems in the peer review process carried out at a distance ("remote evaluation" or Taylorised assessment), often used to minimise costly interaction among experts.

However, there is no reason to think that the VQR peer review process has been vulnerable to such problems. The involvement of many external experts, each entrusted with a reasonable number of 'research outcomes' submitted for evaluation, minimised any concerns that an excessive workload would impinge on the ability to read the publications carefully and assess them in a balanced way.

One may argue that the two evaluations were not independent since a similar exercise had preceded the rating of journals for ASN purposes. This had been implemented by the expert panels that were supervising the VQR process (see Baccini

2016). However, this last exercise had a much more limited scope and identified only a relatively small number of top journals, starting from pre-determined lists which were provided by scientific societies. The primary goal of journal rating within the VQR was to identify those journals from which it was expected that Italian researchers would submit their *best* articles. It was a top-down exercise. On the contrary, the ASN exercise started from the total number of entries in the self-administered website of scholars, which included more than 60,000 entries. It was a bottom-up exercise and led to a much larger number of classifications. The smaller number of journals in the VQR exercise was arranged into up to three tiers, while the ASN journal ratings (devised by a smaller number of experts, who were *not* members of the VQR panels) were derived from parsing a much larger set of publication venues to exclude those which were non-academic. Academic journals were arranged in only two tiers – A-class versus the rest.

The top tiers of the two lists overlap to some extent. This mostly results from reputational factors – at least a large part of the affected academic communities share the opinion that some academic journals (because of their history, intellectual origin, editorial policy, editorial board reputation and scientific committee) possess higher qualities than others. Once we account for the influence of these reputational aspects, it is fair to conclude that the two exercises of journal ratings discussed here were different enough in their purposes and methods as to be considered as substantially, if not entirely, independent. As noted above, the paper by Bertocchi et al. (2015) shows that, in economic disciplines, scores given by referees in VQR to individual articles (our dependent measure in the model below) did not automatically follow the rating of the journals in which these articles were published. A similar remark can be found in the case of legal studies. An essential element to exclude that referees followed the VQR journal rating in evaluating articles automatically so that our model's variables were not independent.

It is possible to carry out a controlled experiment, extending to all SSH, except economics and business, using the analysis initiated by Ferrara and Bonaccorsi (2016) for journals of anthropology, education sciences, geography, history, library science, palaeography, and philosophy. In the following, we first introduce the database used for the analysis and test the influence of the journal classification on the article score. The paper will be concluded with a discussion of the results obtained.

4 Methodology

The paper is based on a dataset including data on all the journal articles submitted for evaluation by Italian scholars in the disciplinary areas of architecture, arts and humanities, history and philosophy, law and sociology and political science. Submissions for evaluation took place within the framework of VQR 2004–10. Italy's national research assessment exercise involved all professors and researchers affiliated to the Italian Universities and Public Research Organisations (PROs) as of November 2011. According to adopted rules, SSH research evaluation was entirely

Table 1 Description of dataset

Area of assessment	Acronym	Full professor	Associate professor	Researcher	Other	N° of articles	N° of articles in Class A journals
Architecture	Area08	280	278	353	7	918	360
Antiquities, philology, literary studies, art history	Area10	1040	1191	1322	26	3579	1954
History, philosophy, pedagogy, and psychology	Area11	713	680	726	8	2127	1086
Law	Area12	1488	983	1337	30	3838	2637
Political and social sciences	Area14	338	409	442	9	1198	664
Total		3859	3541	4180	80	11,660	6701

based on peer review, with the exception of economics and psychology (see the chapters by Bonaccorsi in this volume). Research quality was assessed against *relevance* criteria, intended as contribution to the advancement of the state of art in the field, adequacy, efficacy, timeliness and duration of impacts; *originality and innovation*, intended as a contribution to the creation of new knowledge in the field; *internationalisation*, intended as its position in the international research. Five Groups of Evaluation Experts (GEV in the Italian acronym) carried out the evaluation; one for each SSH area (Architecture; Arts and Humanities; History and Philosophy; Law; Sociology and political sciences). Reviewers were instructed by the GEV to evaluate articles only on the basis or their merit, regardless of the journal in which they were published and of the publication language. Each article had a possible rating of Excellent (A), Good (B), Fair (C) or Limited (D); to each class corresponded to a score ranging from 1 (for articles A-rated) to zero (for articles deemed as limited). Negative scores were assigned if the article was deemed as non-academic (−1) or for plagiarism or fraud (−2, see Ancaiani et al. 2015 for details). In the human and social sciences fields, a substantial fraction of articles – namely, 6701 out of 11,660 (Table 1) – appeared in journals deemed as 'A-class' according to the ASN procedure which was intended to select the best researchers for the ranks of associate and full professors.

According to the relevant Ministerial Decree (No. 76/2012), those journals, were 'internationally recognised as excellent because of the rigor of their peer review procedures and because of their diffusion among, esteem by, and impact on, the scholarly community of a field, as indicated by their presence in the major national and international databases' (our translation). Most of the remaining articles appeared in journals considered as 'academic' for ASN purposes, while a minority were published in journals that remained 'uncategorised'. The dataset's main feature

Table 2 Preliminary analysis of the association between the evaluation of research product and the journal evaluation

		Evaluation of Journal			
		A	Not A	Not academic	Total
Evaluation of research product	A	1344	573	20	1937
	B	3184	1743	92	5019
	C	1322	1096	80	2498
	D	837	1176	150	2163
	Non-academic and other	14	21	8	43
	Total	6701	4609	350	11,660

Pearson $\chi^2 = 630.9$*; p-value = 0.000*

is that it allows a comparison between the evaluations of journals and individual articles.

A preliminary analysis shows that there is a relationship between the evaluation of individual articles and journals where the article is published (Table 2). The non-parametric statistic for categorical data (Pearson χ^2) is statistically significant at 1%,[1] showing that the two distributions are dependent and the two ratings are mutually related. In the following, we will examine this relationship and the controls of several author-level and article-level variables, more thoroughly.

5 The Influence of Journal Classification on the Article Score

We assume that the probability of an article i, published in the journal j, receiving a score equal to $x \in \{-2; 1\}$ is influenced by the class assigned to the journal and the controls for several article characteristics:

$$P\left(Score_{i,j} = x\right) = F\left(Journal\,class_{i,j}, j, Paper\,characteristics_{i,j}\right) \quad (1)$$

Among the controls, we considered the publication language (Italian or not) and the age (distinguishing among three age classes; less than 40 years, between 41 and 55 years and more than 55 years); scientific activity sector (Scientific Areas 8, 10, 11, 12, 14); academic status (full professor; associate professor; researcher; other); and the researcher's gender. We add two binary variables controlling the existence of international co-author(s) and for the referees' nationality (allowing for the possibility of international referees). Finally, we added a variable for the size of the author's scientific area. It uses an ordered probit model, which is an extension of the standard binary probit model, used when the dependent variable has ranked and

[1] All the statistical analyses have been performed using the software STATA ver. 13 (http://www.stata.com/stata13/)

multiple discrete variables, alternatively considering the whole sample or each scientific area. In the first case, we also control the possible area-specific effects.

To avoid the "dummy trap", we normalised those articles written in Italian with no international co-author, evaluated by an Italian reviewer, presented by a female researcher in sociology and political science, aged less than 40. This means that the statistical significance, sign and magnitude of estimated parameters are interpreted as control group differentials. The total available observations amount to 11,660 varying from a minimum of 918 in Architecture to a maximum of 3838 in Law (Table 3).

The main result was that at the aggregate level and in each scientific area the article score was higher as the journal ranking improved. The probability of receiving a high score grew if the article was published in a high-ranking journal according to ASN's experts' evaluation. When assessing the control variables, we confirmed most of the results which already emerged in a previous paper using the same data (Cicero et al. 2016): article scores are higher for papers not written in Italian, with international co-authors, published by an under-40, male, full or associate professor. We found that at an aggregate level, and in most areas, an international reviewer and a lower number of professors in the specific scientific sector (SSD, *Settore Scientifico Disciplinare*) results in a higher article score. A possible interpretation of the first result is that the expert groups responsible for the evaluation assign international reviewers to more international papers which have a greater probability of receiving a high score – given that the level of internationalisation was one of the VQR evaluation criteria (see Ancaiani et al. 2015). The negative relationship between area size and article score result already emerged in Ferrara and Bonaccorsi (2016) and is now extended to all SSH. A possible interpretation is that small fields may be favoured by a "proximity bias" among authors and reviewers, thus resulting, *ceteris paribus*, in higher article scores.

As a final check, once controlling for the same variables considered in model 1, we concentrated on the probability of receiving an excellent score and related it to the fact that the article is published in a top, A-Class journal:

$$P\left(Score_{i,j} = "E"\right) = F\left(Journal\,class_{i,j} = "\mathrm{A}", Paper\,characteristics_{i,j}\right) \qquad (2)$$

In (2), F is the logistic function, and the model is estimated as a logit, a class of models allowing the prediction of a binary response based on the specified predictors. A desirable feature of the logit model is that the regression coefficients may easily be transformed into an odds ratio, expressing the change in the odds of the occurrence under scrutiny (in our case, the odds for a paper of receiving an 'Excellent' evaluation) due to a small change of a given predictor. In our case, we were particularly interested in the odds associated with the classification of a journal as a top, Class A journal. Estimation results for both the aggregate sample and each scientific area are presented in Table 4.

According to logit estimations, the probability of receiving an excellent evaluation is positively affected by the journal in which the paper is published. Publishing

Table 3 Ordered probit model (Dependent variable: article score)

Variables	Total	Architecture	Arts&Hum.	Hist.& Phil.	Law	Sociology & Pol. Sci.
Journal rating	0.417***	0.542***	0.400***	0.379***	0.503***	0.323***
Architecture	0.134***					
Arts and Humanities	0.720***					
Hist. & Philosophy	0.471***					
Law	0.259***					
Italian language	−0.372***	−0.518***	−0.148***	−0.623***	−0.281***	−0.704***
41–55 years	−0.151***	−0.256	−0.208***	0.0492	−0.265***	−0.265**
More than 55 years	−0.582***	−0.662***	−0.726***	−0.394***	−0.563***	−0.572***
Associate professor	0.318***	0.206**	0.308***	0.265***	0.439***	0.234***
Full professor	0.818***	0.788***	0.690***	0.660***	1.096***	0.679***
Other personnel	−0.277**	−0.157	−0.329	0.359	−0.415**	−0.742*
Male	0.0777***	0.184**	0.0882**	0.0257	0.0542	−0.00491
International coauthors	0.301***	0.505***	0.122	0.237	0.902***	0.205
International reviewer	0.153***	0.363***	0.185***	0.0243	0.225***	0.0722
Number of Professors in the scientific sector (SSD)	−0.00131***	−0.00297***	−0.00248***	−0.00146**	−0.000921***	−0.00390***
Constant cut1	−2.854***	−2.514***	−2.081***	−2.990***	−1.867***	−2.953***
Constant cut2	−1.718***	0.363	−0.454***	−2.106***	−1.777***	−0.490***
Constant cut3	−1.695***	1.076***	0.200*	−0.357**	0.378***	0.402**
Constant cut4	0.302***	2.387***	1.574***	0.318**	1.162***	1.539***
Constant cut5	1.026***			1.651***	2.710***	
Constant cut6	2.400***					
Observations	11,660	918	3579	2127	3838	1198
Pseudo R-squared	0.0814	0.0919	0.0543	0.0720	0.0926	0.0832

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4 Logit model (Odds ratio)

Variables	Total	Architecture	Arts, & Hum.	Hist. & Phil.	Law	Sociology & Pol. Sci.
Top journal classification	1.952***	2.513***	1.834***	2.424***	1.990***	1.311
Architecture	1.210					
Arts and Humanities	3.042***					
History and Philosophy	2.031***					
Law	1.084					
Italian language	0.488***	0.311***	0.681***	0.333***	0.529***	0.243***
41–55 years	0.878	0.408**	0.697**	1.144	0.807	0.671
More than 55 years	0.411***	0.248***	0.303***	0.506**	0.572***	0.252***
Associate professor	1.793***	1.283	1.815***	1.825***	2.629***	1.620*
Full professor	4.650***	3.263***	3.831***	4.023***	9.909***	4.877***
Other personnel	1.660	–	1.360	3.057	2.470	1.293
Male	1.247***	1.664**	1.263***	1.155	1.077	1.028
International coauthors	1.611***	2.357**	1.118	1.558	5.149***	1.511
International reviewer	1.352***	1.566**	1.393***	1.178	1.560***	1.490**
Full prof. in the SSD	0.998**	0.992**	0.996**	0.998	1.000	0.990***
Constant	0.065***	0.201***	0.258***	0.137***	0.0332***	0.236***
Observations	11,660	911	3579	2127	3838	1198
Pseudo R-squared	0.116	0.140	0.0738	0.122	0.129	0.143

*** $p < 0.01$, ** $p < 0.05$, *$p < 0.1$

in a Class A journal almost doubled the probability of receiving an excellent evaluation. In each scientific area, the odds of receiving an excellent evaluation were more than doubled by the publication in a Class A journal in architecture and history and philosophy. The effect is somewhat lower, but still highly significant, in law, and arts and humanities, while disappearing in sociology and political sciences. Logit estimation broadly confirmed the results already emerging from the ordered probit model. The odds of receiving an excellent evaluation were increased by publishing in a foreign language, with an international co-author (albeit only in law and architecture) and when the submitting author was 40 years old or younger, an associate or full professor and a male. Gender is significant at the aggregate level and in architecture and humanities, but not in the remaining areas. Having an international

reviewer and publishing in an SSD with a lower number of full professors helped in obtaining an excellent evaluation.

6 Conclusions

Using a large dataset of journal articles published in SSH, the paper proves that independent classifications of journals may be considered as good predictors of the score assigned to individual articles. More specifically, we found that, after controlling some articles' characteristics, the probability of receiving a better score grew with the quality profile of the journal in which the article was published. The probability of receiving an excellent score almost doubles when the paper was published in a top, A-Class journal. The findings held both at the aggregate level and for each specific sub-areas that were considered in the analysis. While peer review must remain the main evaluation methodology, our results indicate that expert-based journal classification may be considered a useful supporting tool in a large evaluation exercise, since it may provide reviewers with valuable information which is apt to support expert evaluation.

References

Aagaard, K., Bloch, C., & Schneider, J. W. (2015). Impacts of performance-based research funding systems: The case of the Norwegian publication indicator. *Research Evaluation, 24*(2), 106–117.

Ahlgren, P., Colliander, C., & Persson, O. (2012). Field normalized citation rates, field normalized journal impact and Norwegian weights for allocation of university research funds. *Scientometrics, 92*(3), 767–780.

Ancaiani, A., Anfossi, A., 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 2004-2010 research evaluation exercise. *Research Evaluation, 24*(3), 242–255.

Archambault, E., Vignola-Gagnè, E., Cotè, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics, 68*(3), 329–342.

Baccini A. (2016). Comments on Bonaccorsi A, Cicero T, Ferrara A. and Malgarini M., *Journal ratings as predictors of articles quality in Arts, Humanities and Social Sciences: an analysis based on the Italian Research Evaluation Exercise*. https://f1000research.com/articles/4-196/v1.

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.

Bolton, K., & Kuteeva, M. (2012). English as an academic language at a Swedish university: Parallel language use and the 'threat of English'. *Journal of Multilingual and Multicultural Development, 33*(5), 429–447.

Butler, L. (2003a). Explaining Australia's increased share of ISI publications – The effects of a funding formula based on publication counts. *Research Policy, 32*, 143–155.

Butler, L. (2003b). Modifying publication practices in response to funding formulas. *Research Evaluation, 12*, 39–46.

Campanario, J.-M. (1998a). Peer review for journals as it stands today – Part 1. *Science Communication, 19*(3), 181–211.

Campanario, J.-M. (1998b). Peer review for journals as it stands today – Part 2. *Science Communication, 19*(4), 277–306.

Campbell, K., Goodacre, A., & Little, G. (2006). Ranking of United Kingdom law journals: An analysis of the research assessment exercise 2001 submissions and results. *Journal of Law and Society, 33*, 335–363.

Carayol, N., & Dalle, J. M. (2007). Sequential problem choice and the reward system in Open Science. *Structural Change and Economic Dynamics, 17*, 167–191.

Christenson, J. A., & Sigelman, L. (1985). Accrediting knowledge: Journal stature and citation impact in social science. *Social Science Quarterly, 66*(4), 964–975.

Chubin, D. E., & Hackett, E. J. (1990). *Peerless science: Peer review and U.S. science policy*. Albany: State University of New York Press.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioural and Brain Sciences, 14*, 119–186.

Cicero, T., Malgarini, M., & Benedetto, S. (2016). Determinants of research quality in Italian universities: Evidence from the 2004 to 2010 evaluation exercise. *Research Evaluation, 25*(3), 257–263.

Cole, S. (1992). *Making science. Between nature and society*. Cambridge, MA: Harvard University Press.

Cole, S., Cole, J. R., & Simon, G. (1981). Chance and consensus in peer review. *Science, 214*, 881–886.

Daniel, H.D. (1993). *Guardians of science. Fairness and reliability of peer review*. Weinheim: VCH.

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.

Finkenstaedt, T. (1990). Measuring research performance in the humanities. *Scientometrics, 19*(5/6), 409–417.

Frost, C. O. (1979). The use of citations in literary research: A preliminary classification of citation functions. *Library Quarterly, 49*(4), 399–414.

Gimenez-Toledo, E., & Roman-Roman, A. (2009). Assessment of humanities and social sciences monographs through their publishers: A review and a study towards a model of evaluation. *Research Evaluation, 18*, 201–213.

Glanzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and the social sciences. *Information Processing and Management, 35*, 31–44.

Gläser, J., & Laudel, G. (2005). Advantages and dangers of 'remote' peer evaluation. *Research Evaluation, 11*(2), 141–154.

Gomez-Mejia, L., & Balkin, D. (1992). Determinants of faculty pay: An agency theory perspective. *Academy of Management Journal, 35*(5), 921–955.

Guetzkow, J., Lamont, M., & Mallard, G. (2004). What is originality in the humanities and the social sciences? *American Sociological Review, 69*(2), 190–212.

Hammarfelt, B. (2011). Interdisciplinarity and the intellectual base of literature studies: A citation analysis of highly cited monographs. *Scientometrics, 86*(3), 705–725.

Hammarfelt, B. (2012). Harvesting footnotes in a rural field: Citation patterns in Swedish literary studies. *Journal of Documentation, 68*(4), 536–558.

Hammarfelt, B. (2016). Beyond coverage: Toward a bibliometrics for the humanities. In M. Ochsner et al. (Eds.), *Research assessment in the humanities* (pp. 115–132). London: Springer Open 2016.

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.

Hasselback, J. R., Reinstein, A., & Schwan, E. S. (2000). Benchmarks for evaluating the performance of accounting faculty. *Journal of Accounting Education, 18*, 79–97.

Hellqvist, B. (2010). Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology, 61*(2), 310–318.

Hemlin, S. (1996). Social studies of the humanities: A case study of research conditions and performance in ancient history and classical archaeology and English. *Research Evaluation, 6*(1), 53–61.

Hemlin, S., & Gustafsson, M. (1996). Research production in the arts and humanities. A questionnaire study of factors influencing research performance. *Scientometrics, 37*(3), 417–432.

Hogler, R., & Gross, M. A. (2009). Journal rankings and academic research: Two discourses about the quality of faculty work. *Management Communication Quarterly, 23*(1), 107–126.

Hug, S. E., Ochsner, M., & Daniel, H.-D. (2013). Criteria for assessing research quality in the humanities: A Delphi study among scholars of English literature, German literature and art history. *Research Evaluation, 22*(5), 369–383.

Hug, S. E., Ochsner, M., & Daniel, H.-D. (2014). A framework to explore and develop criteria for assessing research quality in the humanities. *International Journal of Education Law and Policy, 10*(1), 55–64.

Ingwersen, P., & Larsen, B. (2014). Influence of a performance indicator on Danish research production and citation impact 2000–12. *Scientometrics, 101*(2), 1325–1344.

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.

Lamp, J. W. (2009). At the sharp end: Journal ranking and the dreams of academics. *Online Information Review, 33*(4), 827–830.

Lee, F. S. (2006). The ranking game, class, and scholarship in American mainstream economics. *Journal of Economics, 3*, 1–41.

Macdonald, S., & Kam, J. (2008). Quality journals and gamesmanship in management studies. *Management Research News 31(8),*595-606.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(1977), 161–175.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics, 66*(1), 81–100.

Nederhof, A. J., & Noyons, E. C. M. (1992). International comparison of departments' research performance in the humanities. *Journal of the American Society for Information Science, 43*(3), 249–256.

Nederhof, A. J., & Zwaan, R. A. (1991). Quality judgements of journals as indicators of research performance in the humanities and the social and behavioural sciences. *Journal of the American Society for Information Science, 42*(5), 332–340.

Nederhof, A. J., Zwaan, R. A., De Bruin, R. E., & Dekker, P. (1989). Assessing the usefulness of bibliometric indicators for the humanities and the social sciences – A comparative study. *Scientometrics, 15*(5/6), 423–435.

Ochsner, M., Hug, S. E., & Daniel, H.-D. (2012). Indicators for research quality in the humanities: Opportunities and limitations. *Bibliometrie – Praxis und Forschung, 1*, 4.

Ochsner, M., Hug, S. E., & Daniel, H.-D. (2013). Four types of research in the humanities: Setting the stage for research quality criteria in the humanities. *Research Evaluation, 22*(2), 79–92.

Park, S., & Gordon, M. (1996). Publication records and tenure decisions in the field of strategic management. *Strategic Management Journal, 17*(2), 109–128.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinarity. The case of innovation studies in business and management. *Research Policy, 41*(7), 1262–1282.

Rodriguez-Navarro, A. (2009). Sound research, unimportant discoveries: Research, universities, and formal evaluation of research in Spain. *Journal of the American Society for Information Science, 60*(9), 1845–1858.

Schneider, J. W. (2009). An outline of the bibliometric indicator used for performance-based funding of research institutions in Norway. *European Political Science, 8*(3), 364–378.

Sivertsen, G. (2016). Publication-based funding: The Norwegian model. In M. Ochsner et al. (eds.), *Research Assessment in the Humanities*, Springer Open 2016, London, 79–90.

Starbuck, W. H. (2005). How much better are the most prestigious journals? The statistics of academic publication. *Organization Science, 16*, 180–200.

Thompson, J. W. (2002). The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship. *Libri, 52*, 121–136.

Travis, G. D. L., & Collins, H. M. (1991). New light on old boys: Cognitive and institutional particularism in the peer review system. *Science, Technology & Human Values., 16*(3), 322–341.

Van Fleet, D., McWilliams, A., & Siegel, D. (2000). A theoretical and empirical analysis of journal rankings: The case of formal lists. *Journal of Management, 26*(5), 839–861.

Ward, K. (2009). The future of research monographs: An international set of perspectives. *Progress in Human Geography., 33*(1), 101–126.

Williams, P., et al. (2009). The role and future of the monograph in arts and humanities research. *ASLIB Proceedings, 61*, 67–82.