Scientometrics (2020) 125:533–549
https://doi.org/10.1007/s11192-020-03643-x

Are Italian research assessment exercises size-biased?

Camil Demetrescu[1] · **Andrea Ribichini**[1] · **Marco Schaerf**[1]

Received: 10 February 2020 / Published online: 5 August 2020

Abstract
Research assessment exercises have enjoyed ever-increasing popularity in many countries in recent years, both as a method to guide public funds allocation and as a validation tool for adopted research support policies. Italy's most recently completed evaluation effort (VQR 2011–14) required each university to submit to the Ministry for Education, University, and Research (MIUR) 2 research products per author (3 in the case of other research institutions), chosen in such a way that the same product is not assigned to two authors belonging to the same institution. This constraint suggests that larger institutions, where collaborations among colleagues may be more frequent, could suffer a size-related bias in their evaluation scores. To validate our claim, we investigate the outcome of artificially splitting Sapienza University of Rome, one of the largest universities in Europe, in a number of separate partitions, according to several criteria, noting significant score increases for several partitioning scenarios.

Keywords Research assessment · Bibliometrics · National evaluations · Graph partitioning

Introduction

Research assessment exercises have been adopted by an increasing number of countries in recent years. Their objectives include guiding public funding of research institutions, stimulating improvement through competition and assessing the effectiveness of adopted support policies (Abramo and D'Angelo 2015). Research assessments are conducted through a variety of methodologies, and techniques used in a given country may even vary from one iteration to the next, based on experience, theoretical advancements, availability of resources, and policy aims (Abramo et al. 2011).

✉ Marco Schaerf
marco.schaerf@uniroma1.it

Camil Demetrescu
demetres@diag.uniroma1.it

Andrea Ribichini
ribichini@diag.uniroma1.it

[1] Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Via Ariosto 25, 00185 Rome, Italy

Italy's most recently completed exercise, namely VQR (Research Quality eValuation) 2011–2014, was based on a hybrid approach (i.e., bibliometric indicators for hard sciences and peer review for social sciences and humanities) and examined a relatively small selection of research products (2 per researcher for universities, and 3 per researcher for other institutions, chosen in such a way that no two researchers belonging to the same institution could be assigned the same research product). We refer the reader to "The Italian research evaluation exercises" section for more details. The constraint that researchers from the same institution are not allowed to select the same research products for evaluation might penalize larger institutions, where collaborations among colleagues may be more frequent.

While most authors agree that there is a critical mass of researchers to produce high-quality research (see, e.g Morgan 2004; Adams and Thomson 2011; Kenna and Berche 2011; Calabrese et al. 2018), there is no agreement on whether the size of a university influences the quality and quantity of research publications. Investigations about higher education in the US, for example, have yielded contrasting results, as far as return to size is concerned (Jordan et al. 1988, 1989; Golden and Carstensen 1992). In the same country, however, economies of both scale and scope seem to be at play in the educational market (Koshal and Koshal 1999; Laband and Lentz 2003), even though some authors arrive at different conclusions (Adams and Griliches 1998). Economies of scale have also been detected in several other countries (Hashimoto and Cohn 1997; Avkiran 2001; Izadi et al. 2002; Abbott and Doucouliagos 2003; Longlong et al. 2009; Nemoto and Furumatsu 2014). Once again, however, some authors present contrasting results (Worthington and Higgs 2011). On the other hand, no evidence of size and agglomeration effects have been found in public research institutions such as the Italian National Research Council (CNR) and the French INSERM (Bonaccorsi and Daraio 2005). Moreover, recent articles have revealed constant return to size, and constant return to scope, for research productivity in Italian universities (Abramo et al. 2012, 2014).

We want to investigate whether using the VQR 2011–2014 rules puts larger universities at a disadvantage with respect to small and midsize ones.

Our contribution

We ran simulations in which Sapienza University of Rome (one of the largest universities in Europe) was artificially split into two or more partitions, according to various criteria, computing VQR scores separately for each partition. Our results show marked increases in overall score (i.e., the sum of all partitions' scores) for several partitioning scenarios and we show the impact of this increase in terms of funding and ranking. Large universities may indeed have been penalized by the methodology employed in the most recent national research assessment exercise.

The Italian research evaluation exercises

The history of Italian research assessment exercises begins with VTR (Triennial Research Evaluation) 2001–2003. This evaluation effort required each research institution to submit a number of research products (e.g., publications, patents, etc.) that amounts to 25% of its research staff. Such a small sample size was at least partially justified by the fact that VTR 2001–2003 was based on a pure peer-review process.

VTR 2001–2003 has so far been followed by two more research evaluation exercises, namely VQR (Research Quality eValuation) 2004–2010 and VQR 2011–2014, which closely resemble one another. Both employed a hybrid approach, in which the so-called *bibliometric areas* (e.g., hard sciences) were primarily analyzed through bibliometric indicators, while *non-bibliometric areas* (e.g., social sciences and humanities) underwent a peer-review process. Research products were classified into 16 research areas (numbered from 01 to 14, with areas 08 and 11 split in two parts 08a, 08b, 11a, 11b). Each of the 16 areas was administered by a committee called GEV (Groups of evaluation experts, Gruppo Esperti della Valutazione in Italian).

On both occasions, the VQR program was articulated in two phases. During Phase 1, based on the authors' self-evaluations and guidelines provided by ANVUR (a research evaluation agency instituted by the Italian Ministry of Education, Universities, and Research), each institution selected and submitted (at most) the required number of research products for each one of its authors, in such a way that each product was formally associated with exactly one author. The number of products required for each author varied according to the type of research institution. The default value was 2 for universities and 3 for other research structures for VQR 2011–2014 (3 and 6, respectively, for VQR 2004–2010, which extended over a longer period). During Phase 2 ANVUR formulated its independent quality judgment about the submitted research products (the score assigned to each product is currently revealed only to its authors). The sum of the scores resulting from ANVUR's evaluation was then taken as the VQR score for that research institution.

Both VQRs used a combination of citation counts and journal impact factors (albeit with different combination rules), as they were derived from international databases such as Scopus and WoS, to rate articles in bibliometric areas, resorting to a process of *informed peer review* only in cases of significant discrepancies between the two measures (e.g., highly cited articles in poorly-ranked journals, or vice versa).

A strict requirement of both VQRs is the constraint that each submitted research product must be associated with exactly one of the authors and a university cannot associate the same product to more than one researcher. However, if a paper is co-authored by researchers in different universities, all the participating universities can submit the same product.

The results of the ANVUR evaluation was only made accessible to the authors of the research product, hence we have access to the ANVUR scores only in the aggregated format of the VQR 2011–2014 final report[1] and tables.[2] As the Italian version of the report[3] is more complete, in this article we will refer to the data contained in this version. In particular, this version contains a report for each one of the 16 GEVs, hence, we will use these area reports to assess the reputation impact of the size-bias.

Several authors have analyzed VQR 2011–2014 in depth and raised methodological concerns. Franceschini and Maisano (2017b) highlight that the number of products evaluated for each author is too small to allow identification of excellent, or even average institutions and all that can be detected are the less virtuous ones. Furthermore, they object to the use of journal metrics as a component for article evaluation, as the number of citations received by articles published in the same journal may greatly vary. Another objection concerns the hybrid approach that leads to the combination of peer review and bibliometric analysis, as there is no adequate empirical evidence that they are mutually compatible.

[1] https://www.anvur.it/wp-content/uploads/2017/06/VQR2011-2014_Final%20Report.pdf.
[2] https://www.anvur.it/wp-content/uploads/2017/06/VQR2011-2014_Final%20Report~.rar.
[3] https://www.anvur.it/rapporto-2016/.

Both Franceschini and Maisano (2017b) and Abramo and D'Angelo (2016) argue that the combination of citational data and journal metrics in the evaluation of research products in VQR 2011–2014 is not justified based on current scientific literature. The article by Franceschini and Maisano has triggered a reply by some members of the ANVUR team in charge of the evaluation (Sergio et al. 2017a), followed by response comments from the original authors (Franceschini and Maisano 2017a), an intervention by Abramo and D'Angelo (2017) and further comments by members of the ANVUR team (Sergio et al. 2017b). In essence, the ANVUR team point out the complexities and subtleties of mounting a comprehensive, nation-wide evaluation of research institutions, defending their choices as far as bibliometric indicators and assesment methodologies are concerned, while authors critical of the VQR stand by their objections.

Methodological issues regarding the criteria employed for bibliometric areas have also been identified in the context of VQR 2004–2010 (Abramo and D'Angelo 2015), many of them closely related to those raised concerning VQR 2011–2014, which is hardly surprising considering the similarities between the two evaluation exercises. Ancaiani et al. (2015), on the other hand, provide a detailed and supportive presentation of the evaluation criteria adopted in VQR 2004–2010, also presenting the most relevant conclusions that can be drawn from this exercise. Research quality is reported to be usually higher in certain areas of the country, while no size or age effect seem to emerge; scientific specialization doesn't seem to play a role either.

In order to validate the mixed approach of peer review and bibliometric assessment introduced by VQR 2004–2010, a representative sample of scientific articles in this exercise was experimentally submitted for both evaluations. Ancaiani et al. (2015) report evidence of a significant degree of concordance between the two methods, a conclusion which is however challenged in Baccini and De Nicolao (2017), and defended in Sergio et al. (2017c).

Graziella et al. (2015), in particular, focus on the comparison between informed peer review and bibliometric evaluation in one specific VQR research area, namely Economics and Statistics, in which they report a particularly high agreement between the two processes. This conclusion, however, in disputed by Baccini and De Nicolao (2016a), who impute the higher agreement to the different experimental protocol adopted for Economics and Statistics, when compared to all other VQR research areas. This triggered a response by Bertocchi et al. (2016) and further replies by Baccini and De Nicolao (2016b), with each side remaining on its positions.

Italian research evaluation exercises have been put into an international perspective by Rebora and Turri (2013), who provide a comparison between Italian VTR (and VQR) with their British counterparts (RAE and REF). The authors point out how Italy's first research assessment exercise, VTR 2001–2003, which was conducted 15 years later than the earliest RAE, was inspired by it, while its successors, VQR 2004–2010 and VQR 2011–2014, diverged from their British counterparts due to the use of bibliometric indicators for many scientific areas. It is further claimed that, compared to the UK, research assessment exercises in Italy have been in general characterized by less debate and more passive reception. In fact, part of the discussion on REF involves the possible adoption of metrics, whether based on bibliometric indicators or on altmetrics, to supplement peer review (Stuart 2015) and the reliability and calibration of the peer review process itself (Tymms and Higgins 2018), highlighting concerns not too distant from those surrounding the Italian research evaluation process. Under heavy scrutiny is also the increasing relevance attributed to research impact outside of academia, as an evaluation parameter next to scientific output and research environment (Sutton 2020; Pinar and Unlu 2020). It is

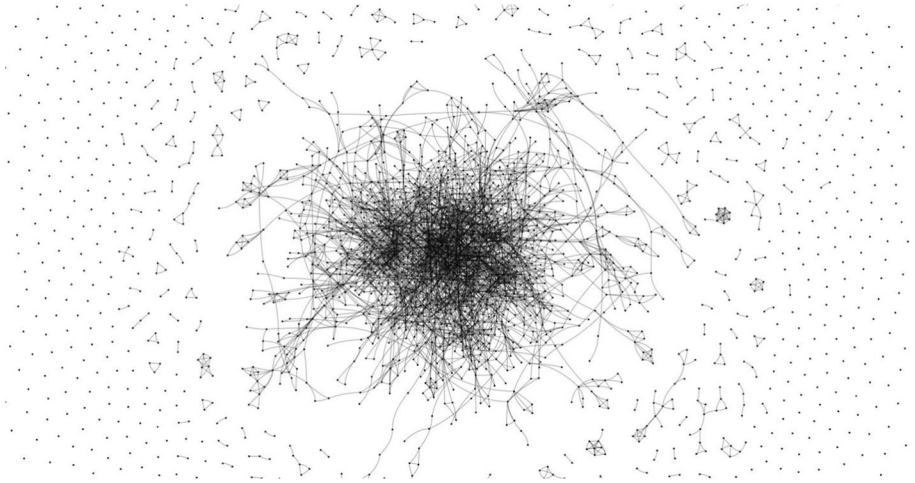

Fig. 1 Coautorship in Sapienza publications submitted to VQR. Each point represents a Sapienza researcher and lines represent coauthorship in at least 1 submitted publication, from Demetrescu et al. (2019)

argued that its adoption may limit scientific freedom, at least in certain research areas, and may also favor larger institutions.

Data and methods

The authors of this work coordinated the participation of Sapienza University of Rome to both VQR 2004–2010 and VQR 2011–2014, thus gaining a deep understanding of the correlations between the various research areas within Sapienza. Our starting point was the database of Sapienza authors that participated in the most recently completed Italian research evaluation exercise (VQR 2011–2014) and the submitted research products associated with each one of them. Overall, there were 3562 authors and 5909 research products submitted to ANVUR for evaluation. In Fig. 1, we show the interconnections between the Sapienza researchers, each researcher is represented by a dot and a link connecting them indicates that they coauthored at least one paper among the set of papers submitted by Sapienza for evaluation in the VQR 2011–2014 assessment exercise. The figure shows that there is a very large kernel of researchers who have many papers coauthored by other Sapienza researchers. As it turns out, the densest web of interconnections is in the life sciences area.

As pointed out in "The Italian research evaluation exercises" section, ANVUR's official scores were only made available to products' authors. However, by applying ANVUR's guidelines and by retrieving bibliometric indicators from Scopus and WoS, we were able to reasonably estimate product scores, at least for bibliometric areas (we refer the interested reader to Demetrescu et al. (2019) for more details). Our estimates tend to underestimate the final result since we cautiously assigned a grade of 0 to all papers which went to peer-review. For this reason, we will assess the impact by computing the percentage increase in our estimate and assume that the same increase would apply to the final score. We suspect that, since our estimates assign a score of 0 to a large number of products, the real impact of the VQR rules is probably larger than what we estimate in this article.

We estimate the impact by creating a scenario where Sapienza is divided in several institutions, each one with a fraction its members. Our simulations include partitioning Sapienza in 1 (no partition), 2, 3 and 4 smaller institutions. We considered partitioning up to 4 since many medium-sized Italian universities are approximately 1/4 of the size of Sapienza. We denote with Sapienza/2, Sapienza/3, and Sapienza/4 a hypothetical university whose size is half, one third or one-fourth of the size of Sapienza. These fictitious universities are then inserted in the list of Italian Universities sorted by size (number of expected research products) in Table 1. Notice that even Sapienza/4 would still be an upper mid-size university. The table also includes the percentage of missing products for each university, which is the difference between the number of expected products with respect to the number of submitted ones. This factor should account for the number of inactive researchers, but in this case, this was influenced by a boycott of the VQR 2011–2014 by many researchers. Notice that Sapienza was one of the universities where the boycott had especially high participation among faculty members, hence its results have been lower than expected due to this large number of missing products.

Results

In this section, we compute the penalty that has occurred to Sapienza due to its size and the VQR rules. To understand why these rules had a significant impact on the score obtained by Sapienza, we point out that there is a very significant amount of collaboration between researchers, even when they belong to different areas of research, as clearly shown by Fig. 1.

To more precisely compute the impact of the no-duplicates rule, we would need to have the actual grades assigned by VQR to any publication. Unfortunately, this data is not publicly available, so we resort to an estimate of the grades that we were able to compute for the areas that have been evaluated using mainly bibliometric criteria. Therefore, we restrict our analysis to the bibliometric areas and we exclude authors from non-bibliometric areas, With this restriction, the number of Sapienza researchers evaluated under VQR amounts to 2816.

Let R be the set of researchers, P the set of submitted products, n the number of partitions and $score(R, P)$ be the maximum score that can be obtained by the set R using only products in P, by R_i we denote the set of researchers in partition i. We define the percentage increase inc in the score as:

$$inc = 100 \times \left(\frac{(\sum_{i=1}^{n} score(R_i, P)) - score(R, P)}{score(R, P)} \right)$$

Notice that all scores are computed on the same set P of products since a researcher in a partition can now use a product even if it is used by a researcher in another partiton. The score function can be efficiently computed, see Demetrescu et al. (2019) for more details.

There is an obvious upper bound to the percentage increase and it is $(n - 1) \times 100$, but this is highly unrealistic. A much more realistic upper bound can be computed by partitioning the set R in partitions of 1 researcher each. This can be computed easily and we get, in our case, an overall score increase of approximately 12.31%. It is clear that this is an upper bound to any other partitioning scheme.

Table 1 Italian Universities sorted by decreasing value of the expected number of products

University	Expected #	Missing %	University	Expected #	Missing %
Roma La Sapienza	6861	13.2	Napoli Parthenope	593	7.9
Bologna	5095	2.9	Cassino	588	4.1
Napoli Federico II	4504	6.0	Tuscia	568	1.2
Padova	3892	3.9	Macerata	549	1.5
Milano	3780	4.6	Camerino	542	6.1
Torino	3674	2.9	Molise	516	16.9
Sapienza/2	3430	13.2	Reggio Calabria	515	6.8
Firenze	3127	2.8	Bari Politecnico	512	0.4
Palermo	2968	3.4	Milano Bocconi	511	9.0
Bari	2742	6.6	Catanzaro	462	4.5
Pisa	2673	4.6	Teramo	427	2.6
Roma Tor Vergata	2603	5.4	Napoli L'Orientale	371	6.7
Milano Cattolica	2530	8.1	Sannio	359	6.1
Milano Politecnico	2443	8.0	Venezia Iuav	290	5.9
Catania	2372	13.7	Bolzano	278	3.2
Genova	2366	10.6	Enna Kore	259	2.7
Sapienza/3	2287	13.2	Roma Marconi	247	14.6
Messina	2161	9.5	Milano San Raffaele	231	0.0
Perugia	1975	4.1	Roma LUISS	199	2.0
Napoli II	1843	7.7	Pisa S.Anna	197	0.0
Salerno	1779	2.6	Roma Biomedico	192	1.0
Pavia	1773	5.1	Novedrate e-Campus	187	24.6
Cagliari	1757	9.9	Milano IULM	181	0.6
Sapienza/4	1715	13.2	Napoli Benincasa	177	2.8
Parma	1661	5.0	Roma LUMSA	177	2.8

Table 1 (continued)

University	Expected #	Missing %	University	Expected #	Missing %
Milano Bicocca	1646	1.7	Roma UNINETTUNO	144	36.8
Roma Tre	1621	11.6	Pisa Normale	136	2.9
Calabria (Arcavacata di Rende)	1544	10.8	Trieste SISSA	124	1.6
Torino Politecnico	1491	2.0	Roma Foro Italico	118	1.7
Modena e Reggio Emilia	1465	4.6	Roma UNICUSANO	115	0.9
Siena	1425	4.7	Perugia Stranieri	112	1.8
Verona	1353	6.0	Aosta	97	8.2
Chieti e Pescara	1320	3.7	Roma Europea	92	2.2
Trieste	1257	5.4	Roma Link Campus	90	30.0
Udine	1248	4.5	Castellanza LIUC	78	3.8
Sassari	1235	6.9	Napoli Pegaso	77	1.3
Ferrara	1173	2.6	Siena Stranieri	73	2.7
Salento	1169	14.8	Casamassima LUM	72	2.8
Trento	1072	3.5	Roma UNITELMA	65	6.2
Brescia	1048	3.0	Roma UNINT	57	1.8
L'Aquila	1013	8.3	Roma San Raffaele	49	38.8
Marche	974	2.5	Milano HUMANITAS	46	0.0
Venezia Cà Foscari	963	0.7	Benevento - Giustino Fortunato	42	14.3
Piemonte Orientale	709	3.4	Pavia IUSS	35	0.0
Insubria	700	1.9	Lucca - IMT	29	0.0
Foggia	681	3.5	Bra - Scienze Gastronomiche	26	0.0
Urbino Carlo Bo	647	13.3	Roma Mercatorum	24	0.0
Basilicata	607	12.2	Reggio Calabria—Dante Alighieri	15	0.0
Bergamo	605	1.8			

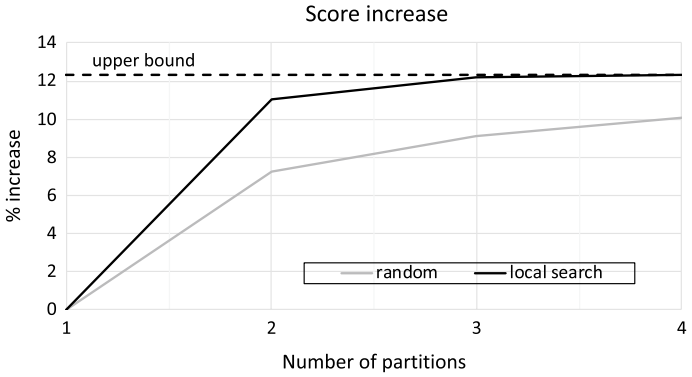

Fig. 2 Partitioning of Sapienza. VQR score percentage increases versus number of partitions using the best algorithm

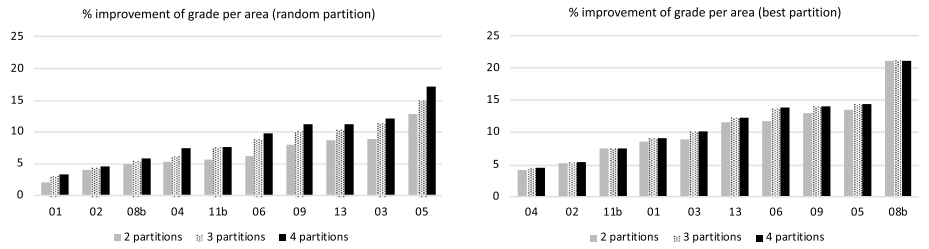

Fig. 3 Percentage of increase of grades per area

To formulate more realistic scenarios, for any value of n, we choose two strategies. In the first one, we randomly split the set R into n subsets and then compute the average increase over 10 trials, while in the second strategy we compute the partitions that attempt to maximize the increase.

The first strategy is easy to implement, the second strategy, however, is more complex. In fact, we have proven that it is not computationally feasible to compute the maximum increase, since the problem of computing the most convenient partitioning is NP-hard, as easily shown by reduction from the well know Simple-Max-Cut, a problem that has been proven to be NP-hard in Karp (1972). Our proof can be found in the "Appendix".

Therefore, we devised a local search heuristics that splits the set R into n subsets while, at the same time, maximizes the overall score and trying to minimize the differences in the partitions. In this case, Sapienza is divided into n random partitions of approximately equal size. At each iteration, the author list is permuted randomly, and then scanned looking for an author whose transfer from one partition to the other would cause the overall score to increase by an amount larger than a set threshold t. When such an author is found, the transfer is enacted, and the heuristic moves on to the next iteration. The algorithm terminates when no partition switch can cause a score increment larger than t.

If we choose $n = 2$ and set t to the minimum possible increase > 0, our local search heuristic converges to an overall score increase of 11.03%, which is fairly close to the

Table 2 Average number of Sapienza coauthors and marks per area with improvements entailed by Sapienza/4 best

Scientific area	% improvement (Sapienza/4 best)	Collaborations	Average mark	SD of marks
01	9.16	1.40	61.08	43.07
02	5.31	2.39	85.50	29.25
03	10.23	2.25	77.94	32.44
04	4.53	1.64	72.11	36.49
05	14.46	2.02	70.90	36.09
06	13.91	2.53	62.29	39.30
08b	21.20	1.53	51.55	42.34
09	14.00	1.65	67.62	40.28
11b	7.426	1.64	55.18	44.85

theoretical maximum of 12.31%. In this case, a fairly straightforward single-threaded implementation takes roughly 8.6 hours to run on our experimental platform. For $n = 4$, we obtain an increase of 12.30%, almost equal to the upper bound.

In the random scenario, when $n = 2$ we obtain an average score increase of 7.26%, for $n = 4$ it reaches 10.05%. The results are summarized in Fig. 2. In Fig. 3 we report for each scientific area and number of partitions the percent of improvement using the best algorithm using the random partitioning scheme (left) and the local search heuristic (right).

The results clearly show that large universities, such as Sapienza, are penalized under the VQR 2011–2014 rules. All considered partitions of Sapienza, whether random or not, will lead to an increase in the score by at least 7% and up to 12.3%.

As it is well known from the literature, different research areas exhibit different propensities for collaboration, both at the extramural and, perhaps more relevant to our case, at the intramural level (see, e.g., Larivière et al. 2006; Abramo et al. 2013a). Bearing this in mind, and with the aim of investigating the reasons behind the observed different levels of improvement across areas, we introduced two additional metrics for research products, namely the average number of internal coauthors, and the average mark. Both metrics were computed per area, and on the whole set of scientific products submitted to ANVUR by Sapienza University. The values obtained, which refer to Sapienza/4 best, are shown in Table 2, where we only concentrated on the areas for which we had statistically significant data. More precisely, we decided to discard area 07 (Agriculture and Veterinary Sciences) where Sapienza is barely active and areas 08b and 13 where the number of bibliometric publications is less than 50%. Next, in order to extract a clearer message from the table, we investigated whether the improvement is correlated to these metrics. We found: (1) that the correlation of improvement with the average number of internal (i.e., Sapienza) authors to be -0.1, (2) that with the average mark to be -0.544, and (3) finally that with the standard deviation of the mark to be 0.358.

We note that there is only a very weak (and inverse) correlation between improvement and average number of coauthors. On the other hand, there is a clear inverse correlation between the average mark and the improvement. The improvement is therefore more relevant in areas where the average mark is lower. This is due to the fact that, in areas where the average mark is already very high, the no-duplicates rule may force the use of less valued products, but their values (marks) are just below the values of the products they

Table 3 Gender analysis

Area	% Female researchers	Collaboration			% improvement		
		Overall	F	M	Overall	F	M
01	35.19	1.40	1.46	1.37	9.16	12.59	7.49
02	10.61	2.40	2.92	2.33	5.31	16.77	4.33
03	47.10	2.25	2.25	2.25	10.23	12.89	7.72
04	32.70	1.63	1.76	1.57	4.53	8.56	2.84
05	53.70	2.02	2.215	1.81	14.46	15.99	12.94
06	36.74	2.53	2.67	2.45	13.91	14.06	13.83
08b	21.94	1.53	1.61	1.51	21.20	14.38	23.24
09	17.88	1.65	1.68	1.64	14.00	6.21	16.11
11b	57.53	1.64	1.76	1.48	7.42	6.96	7.84
13	41.07	1.29	1.35	1.25	12.37	17.83	9.35

substitute. This idea is further supported by the correlation of the improvement with the standard deviation of the marks, which implies that there are increased chances of improvement if the set of marks is more diverse.

Gender impact

In this section we investigate whether the no-duplicates rule has a different impact on female and male researchers in different scientific areas. The results are summarized in Table 3, where we report, for every scientific area, the percentage of female researchers, the average number of Sapienza coauthors in the set of submitted publications (divided into overall, female, and male) and the percentage of improvement in the best 4-way split, again divided into overall, female, and male.

Recent studies (see, e.g., Bozeman and Gaughan 2011; Abramo et al. 2013b, 2019) point out that female researchers have a higher propensity for intramural collaborations than their male colleagues. Table 3 shows that Sapienza University is no exception in this respect. Moreover, the no-duplicates rule has a stronger impact on women than it has on men in all research areas except two, namely Engineering areas 08b and 09, which however have a small percentage of female researchers.

Our data suggests that the VQR rules penalize female researchers in large universities more than their men colleagues.

Impact of VQR scores

Unfortunately, we do not have the necessary data to compute the potential improvement of other universities if the no-duplicates rule were lifted, as data from ANVUR are provided in aggregated form only. As a consequence, in our analysis we assume that no other universities are split, which may result in an overestimation of the rule's impact upon Sapienza University. On the other hand, Sapienza partitions are compared with universities which, undivided, have roughly the same size as those partitions. The goal of this section is to show that the impact of size-penalty might be quite relevant from more than one point of view.

Table 4 VQR funding allocation

Year	Total amount (M€)	Sapienza amount (M€)
2016	1204.025	80
2017	1228.48	86.6
2018	1354.79	86.5

There are at least two potential impacts of the VQR rules on the assessment of Sapienza, first of all, an economic impact, since a relevant part of the Italian university system funding (called FFO in Italian) is directly related to the VQR score, and secondarily an impact on the prestige of the institution, since the rankings have been made public.

Regarding the first aspect, Table 4 summarizes the amount of money that the Italian Ministry for Education, University, and Research (MIUR) has allocated to the universities in the years 2016–2018, according to their VQR score. Over 10% of the total funding of the Italian University System was allocated according to these results.

Over the three years considered (and with the caveat discussed at the start of the section), the money damage incurred by Sapienza can be estimated in the order of 10% of the above amounts, totaling over 25 M€.

The impact on reputation is more difficult to assess, due to the multitude of rankings produced by ANVUR. We decided to concentrate on area rankings, where the impact is more easily assessable. ANVUR produced, for every scientific area (assessed by a GEV) a table with the number of expected products, submitted products, total score and average score for every university which contributed more than 5 products to the area. The GEV reports rank universities according to the average grade obtained, but dividing universities according to their size (in their research area) into 3 classes: small (S), medium (M) or large (L). Sapienza, in almost all research areas, belongs to the class of large universities (one exception is area 07 Agriculture and Veterinary Sciences, in which Sapienza has very few researchers), but the partitioning of Sapienza might classify it as medium or (in a few cases) small-size.

For every research area, we classified Sapienza, and its partitions Sapienza/2 random and Sapienza/4 best, in the appropriate class (S, M, or L) and computed its position. At the same time, we ranked all the universities in a single list and computed the rank of Sapienza split randomly into 2 smaller universities (Sapienza/2 random) and Sapienza split into 4 smaller universities maximizing the total score (Sapienza/4 best). In all cases, for every research area, we also computed the *R* indicator, that is simply the average score of a university in that area divided by the national average for the area. The results are presented in Table 5.

The table clearly shows that the reputation impact can be relevant for a large university such as Sapienza. Ranks improve significantly and, in most cases, when Sapienza's *R* indicator was below average (< 1), it goes above average at least in the case of Sapienza/4 best.

Conclusions

In this article, we have investigated the impact of the no-duplicates rule employed for Italy's most recent research evaluation exercises, which forced each institution to submit a specific number of research products per author, in such a way that each product was

Table 5 Impact on area rankings and R indicator

		Sapienza				Sapienza/2 random				Sapienza/4 best			
Area	#Univ	Overall rank	R	Size	rank in size	Overall rank	R	Size	Rank in size	Overall rank	R	Size	Rank in size
01	59	23	1.04	L	3/7	15	1.1	M	4/19	14	1.14	M	3/19
02	55	25	1.04	L	4/9	14	1.07	M	7/25	9	1.1	M	4/25
03	56	34	0.97	L	6/7	2	1.02	M	5/15	18	1.07	S	15/36
04	43	13	1.08	L	4/9	10	1.1	M	6/18	5	1.14	S	3/18
05	62	46	0.93	L	7/10	32	1.03	L	6/10	21	1.08	M	7/22
06	51	40	0.89	L	9/14	35	0.98	M	13/18	27	1.02	S	16/22
08b	51	38	0.87	L	4/4	29	0.98	S	24/41	18	1.06	S	17/41
09	63	43	0.95	L	6/6	27	1.02	M	3/15	18	1.09	S	19/44
11b	55	15	1.09	L	4/6	12	1.13	L	3/6	9	1.17	M	3/7
13	82	63	0.74	L	6/6	57	0.79	M	30/34	47	0.83	M	28/34

assigned to at most one of its coauthors. More precisely, we have analyzed its impact on Sapienza University, the largest university in Italy and one of the largest in Europe, and argued that this rule may have induced a size-related bias (i.e., larger institutions, due to more frequent collaborations among colleagues, may have been penalized) in the final assessment.

In order to perform our analysis, we have run simulations in which we have artificially split Sapienza University of Rome into several partitions, according to various criteria. We found that in several cases the overall score, given by the sum of all partitions' scores, yielded significant increases over the score obtained by considering Sapienza as a single entity, providing evidence that indeed larger research institutions in Italy may have been penalized by the methodology employed in recent research evaluation exercises. The analysis has also been carried for the various research areas defined in the VQR2004–2010 call and by gender, pointing out that the impact is not uniform across the areas and that there is also a small but relevant gender bias.

Acknowledgements Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

Appendix: NP-hardness proof

In this Appendix we prove that finding the optimal way to partition the set of authors in two subsets such that the overall score is maximized cannot be accomplished in a computationally efficient way. Using the notation of Demetrescu et al. (2019), we define an allocation problem as $A = \langle N, \mathbb{G}, \Omega, \texttt{val}, k\rangle$ comprising a set of agents N and a set of goods $\mathbb{G}$, whose values are given by the function `val` mapping each good to a non-negative real number. The function Ω associates each agent with the set of goods their are interested in. Moreover, the natural number k provides the maximum number of goods that can be assigned to each agent. Each good is indivisible and can be assigned at most to one player.

Theorem 1 *Optimal-partition is NP-hard*

Proof Let $G = (V, E)$ be an undirected graph, we construct an allocation problem $A = \langle N, \mathbb{G}, \Omega, \texttt{val}, k\rangle$ where:

1 $N = V$
2 For any edge $(v_i, v_j) \in E$ there is a good $g_{ij} \in G$, $g_{ij} \in \Omega(v_i)$, and $g_{ij} \in \Omega(v_j)$
3 For any good $g \in G$, $\texttt{val}(g) = 1$
4 $k = 1$

For any partition of A into A_1 and A_2, $\texttt{val}(A_1) + \texttt{val}(A_2) = |E| + k$, where k is the number of edges that connect agents in different partitions. In fact, every good g that is shared between agents in the same partition can be used only once, while when it is shared across the two partitions it can be used for both partitions. Since $|E|$ is constant for any instance, by maximimzing the sum of the values of the two partitions, we are also maximizing the cut in the graph partition. Therefore, the value of a partition of A induces a simple Max-Cut on G. Since simple Max-Cut is NP-hard (see Karp 1972), Optimal-partition is NP-hard □

References

Abbott, M., & Doucouliagos, C. (2003). The efficiency of Australian universities: A data envelopment analysis. *Economics of Education Review*, *22*(1), 89–97. https://doi.org/10.1016/S0272-7757(01)00068-1.

Abramo, G., Cicero, T., & D'Angelo, C. A. (2012). Revisiting size effects in higher education research productivity. *Higher Education*, *63*(6), 701–717.

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*(11), 2202–2214. https://doi.org/10.1002/asi.23323.

Abramo, G., & D'Angelo, C. A. (2016). Refrain from adopting the combination of citation and journal metrics to grade publications, as used in the Italian National Research Assessment Exercise (VQR 2011–2014). *Scientometrics*, *109*(3), 2053–2065. https://doi.org/10.1007/s11192-016-2153-5.

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. https://doi.org/10.1016/j.joi.2017.06.003.

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2011). National research assessment exercises: The effects of changing the rules of the game during the game. *Scientometrics*, *88*(1), 229–238. https://doi.org/10.1007/s11192-011-0373-2.

Abramo, G., D'Angelo, C. A., & Murgia, G. (2013a). The collaboration behaviors of scientists in Italy: A field level analysis. *Journal of Informetrics*, *7*(2), 442–454. https://doi.org/10.1016/j.joi.2013.01.009.

Abramo, G., D'Angelo, C. A., & Murgia, G. (2013b). Gender differences in research collaboration. *Journal of Informetrics*, *7*(4), 811–822. https://doi.org/10.1016/j.joi.2013.07.002.

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2014). Investigating returns to scope of research fields in universities. *Higher Education*, *68*(1), 69–85.

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2019). A gender analysis of top scientists' collaboration behavior: Evidence from Italy. *Scientometrics*, *120*(2), 405–418.

Adams, J., & Thomson, S. (2011). *Funding research excellence: research group size, critical mass & performance*. London: University Alliance. ISBN 9781908190086.

Adams, J. D., & Griliches, Z. (1998). Research productivity in a system of universities. *Annales d'Économie et de Statistique*, *49/50*, 127–162.

Ancaiani, A., Anfossi, A. F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., et al. (2015). Evaluating scientific research in Italy: The 2004–10 research evaluation exercise. *Research Evaluation*, *24*(3), 242–255. https://doi.org/10.1093/reseval/rvv008.

Avkiran, N. K. (2001). Investigating technical and scale efficiencies of Australian Universities through data envelopment analysis. *Socio-Economic Planning Sciences*, *35*(1), 57–80. https://doi.org/10.1016/S0038-0121(00)00010-0.

Baccini, A., & De Nicolao, G. (2016a). Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. *Scientometrics*, *108*(3), 1651–1671. https://doi.org/10.1007/s11192-016-1929-y.

Baccini, A., & De Nicolao, G. (2016b). Reply to the comment of bertocchi. *Scientometrics*, *108*(3), 1675–1684. https://doi.org/10.1007/s11192-016-2055-6.

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*(4), 353–357, 04. https://doi.org/10.1093/reseval/rvx013.

Benedetto, S., Checchi, D., Graziosi, A., & Malgarini, M. (2017a). Comments on the paper "Critical remarks on the Italian assessment exercise", *Journal of Informetrics, 11*(2017), 337–357. *Journal of Informetrics, 11*(2), 622–624 (2017a). https://doi.org/10.1016/j.joi.2017.03.005.

Benedetto, S., Checchi, D., Graziosi, A., & Malgarini, M. (2017b). Comments on the correspondence "On tit for tat: Franceschini and Maisano versus ANVUR regarding the Italian research assessment exercise

VQR 2011–2014". *Journal of Informetrics, 11*(2017), 783–787. *Journal of Informetrics, 11*(3), 838—840 (2017b). https://doi.org/10.1016/j.joi.2017.07.002.

Benedetto, S., Cicero, T., Malgarini, M., & Nappi, C. (2017c). Reply to the letter on Ancaiani et al. "Evaluating scientific research in Italy: The 2004–10 research evaluation exercise". *Research Evaluation, 26*(4), 358–360 (2017c). https://doi.org/10.1093/reseval/rvx017.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., & Peracchi, F. (2015). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Research Policy*, *44*(2), 451–466. https://doi.org/10.1016/j.respol.2014.08.004.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C., & Peracchi, F. (2016). Comment to: Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. *Scientometrics*, *108*, 05. https://doi.org/10.1007/s11192-016-1965-7.

Bonaccorsi, A., & Daraio, C. (2005). Exploring size and agglomeration effects on public research productivity. *Scientometrics*, *63*(1), 87–120.

Bozeman, B., & Gaughan, M. (2011). How do men and women differ in research collaborations? An analysis of the collaborative motives and strategies of academic researchers. *Research Policy*, *40*(10), 1393–1402. https://doi.org/10.1016/j.respol.2011.07.002.

Calabrese, A., Capece, G., Costa, R., Di Pillo, F., & Giuffrida, S. (2018). A 'power law' based method to reduce size-related bias in indicators of knowledge performance: An application to university research assessment. *Journal of Informetrics*, *12*(4), 1263–1281.

Demetrescu, C., Lupia, F., Mendicelli, A., Ribichini, A., Scarcello, F., & Schaerf, M. (2019). On the Shapley value and its application to the Italian VQR research assessment exercise. *Journal of Informetrics*, *13*(1), 87–104. https://doi.org/10.1016/j.joi.2018.11.008.

Franceschini, F., & Maisano, D. (2017a). A rejoinder 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*(3):645–646. https://doi.org/10.1016/j.joi.2017.05.013.

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

Golden, J., & Carstensen, F. V. (1992). Academic research productivity, department size and organization: Further results, comment. *Economics of Education Review*, *11*(2), 153–160.

Hashimoto, K., & Cohn, E. (1997). Economies of scale and scope in Japanese private universities. *Education Economics*, *5*(2), 107–115. https://doi.org/10.1080/09645299700000010.

Izadi, H., Johnes, G., Oskrochi, R., & Crouchley, R. (2002). Stochastic frontier estimation of a CES cost function: The case of higher education in Britain. *Economics of Education Review*, *21*(1), 63–71. https://doi.org/10.1016/S0272-7757(00)00044-3.

Jordan, J. M., Meador, M., & Walters, S. J. K. (1988). Effects of department size and organization on the research productivity of academic economists. *Economics of Education Review*, *7*(2), 251–255. https://doi.org/10.1016/0272-7757(88)90049-0.

Jordan, J. M., Meador, M., & Walters, S. J. K. (1989). Academic research productivity, department size and organization: Further results. *Economics of Education Review*, *8*(4), 345–352. https://doi.org/10.1016/0272-7757(89)90020-4.

Karp, R. M. (1972). Reducibility among combinatorial problems. In R. E. Miller, J. W. Thatcher, & J. D. Bohlinger (Eds.), *Complexity of Computer Computations. The IBM Research Symposia Series*. Boston, MA: Springer. https://doi.org/10.1007/978-1-4684-2001-2_9.

Kenna, R., & Berche, B. (2011). Critical mass and the dependency of research quality on group size. *Scientometrics*, *86*(2), 527–540.

Koshal, R. K., & Koshal, M. (1999). Economies of scale and scope in higher education: A case of comprehensive universities. *Economics of Education Review*, *18*(2), 269–277. https://doi.org/10.1016/S0272-7757(98)00035-1.

Laband, D. N., & Lentz, B. F. (2003). New estimates of economies of scale and scope in higher education. *Southern Economic Journal*, *70*(1), 172–183.

Larivière, V., Gingras, Y., & Archambault, É. (2006). Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. *Scientometrics*, *68*(3), 519–533.

Longlong, H., Fengliang, L., & Weifang, M. (2009). Multi-product total cost functions for higher education: The case of Chinese research universities. *Economics of Education Review*, *28*(4), 505–511. https://doi.org/10.1016/j.econedurev.2008.11.002.

Morgan, K. J. (2004). The research assessment exercise in English universities, 2001. *Higher Education*, *48*(4), 461–482.

Nemoto, J., & Furumatsu, N. (2014). Scale and scope economies of Japanese private universities revisited with an input distance function approach. *Journal of Productivity Analysis*. https://doi.org/10.1007/s11123-013-0378-3.

Pinar, M., & Unlu, E. (2020). Evaluating the potential effect of the increased importance of the impact component in the Research Excellence Framework of the UK. *British Educational Research Journal*, *46*(1), 140–160. https://doi.org/10.1002/berj.3572.

Rebora, G., & Turri, M. (2013). The UK and Italian research assessment exercises face to face. *Research Policy*, *42*(9), 1657–1666. https://doi.org/10.1016/j.respol.2013.06.009.

Stuart, D. (2015). Finding "good enough" metrics for the UK's Research Excellence Framework. *Online Information Review*, *39*(2), 265–269.

Sutton, E. (2020). The increasing significance of impact within the Research Excellence Framework (REF). *Radiography*. https://doi.org/10.1016/j.radi.2020.02.004.

Tymms, P., & Higgins, S. (2018). Judging research papers for research excellence. *Studies in Higher Education*, *43*(9), 1548–1560. https://doi.org/10.1080/03075079.2016.1266609.

Worthington, A. C., & Higgs, H. (2011). Economies of scale and scope in Australian higher education. *Higher Education*, *61*(4), 387–414.