Scientometrics (2026) 131:3699–3716
https://doi.org/10.1007/s11192-026-05656-4

Faculty shuffle: simulating institutional gaming in research evaluation

Giovanni Abramo[1] · **Pasquale Carotenuto**[2] · **Ciriaco Andrea D'Angelo**[3] · **Alessio Salvatore**[4]

Received: 17 April 2025 / Accepted: 30 April 2026 / Published online: 16 May 2026
© The Author(s) 2026

Abstract
This study explores potential institutional 'gaming' behaviors among Italian universities within the context of the VQR 2020–2024 national research evaluation exercise. Using mixed-integer linear programming to simulate potential faculty reallocations across departments, we assess the extent to which strategic internal restructuring can affect university performance outcomes. Results indicate that universities with lower baseline research scores have stronger incentives for internal reorganization. Furthermore, while intra-departmental collaboration provides opportunities for gaming through duplicable outputs, it is the quality of collaborative research that significantly drives evaluation score improvements. The findings highlight critical policy implications, calling for refined evaluation criteria and increased transparency to mitigate gaming behaviors.

Keywords Institutional gaming · Research assessment exercises · Faculty reallocation · Bibliometrics

✉ Ciriaco Andrea D'Angelo
dangelo@dii.uniroma2.it

Giovanni Abramo
giovanni.abramo@unimercatorum.it

Pasquale Carotenuto
p.carotenuto@iac.cnr.it

Alessio Salvatore
alessio.salvatore@cnr.it

[1] Laboratory for Studies in Research Evaluation, Department of Engineering and Science, Universitas Mercatorum, Rome, Italy

[2] Institute for the Applications of Calculus "M. Picone", National Research Council, Rome, Italy

[3] Department of Business Engineering, University of Rome "Tor Vergata", Rome, Italy

[4] Institute for System Analysis and Computer Science "A. Ruberti", National Research Council, Rome, Italy

Introduction

"Institutional gaming" generally refers to the strategic maneuvers, behaviors, or practices within formal organizations or institutions, often with the aim of gaining advantage, influence, or power within the established structures. In scientific literature, it may be discussed in the context of organizational theory, focusing on the dynamics of power, decision-making processes, and the interactions between actors within institutional settings. Specifically, in our study, we refer to how universities organize themselves to undergo national research assessment exercises, on the basis of whose results they will obtain significant shares of block funding necessary for their daily operations and strategic investments.

The term institutional 'gaming' is here used to distinguish it from individual 'gaming' (such as unethical publication and citation practices), which have been more widely studied and dissected in literature (Abramo et al., 2019; Graf et al., 2019; Groen-Xu et al., 2023).

Scholarly effort to investigate institutional 'gaming' is crucial because of its implications for organizational effectiveness, accountability, transparency, and equity, as well as its potential to undermine the intended purposes or public trust in academia and the public research system as a whole. Yet, contributions in the literature are very few indeed, and all focused on the UK research assessment exercises (RAEs) now shifted to REF (Research Excellence Framework). In particular, Stern draws attention to the human resources practices of UK universities hiring staff on "fractional contracts" (Stern, 2016a, 2016b) in response to the RAE/REF rules. Although without any quantitative supporting evidence, he speaks of a "demonstrable increase in the number of individuals being recruited from other institutions shortly before the census date," labeling such practices as "gaming the system." Recently, Harris et al. presented for the first time quantitative evidence of strategic hiring practices adopted by UK universities (in the Russell group) in responding to the RAE/REF cycles (Harris et al., 2025). Using higher education staff population panel data and focusing on the RAE/REF cut-off census date, the authors found strong evidence of a 2008 and 2014 "cycle", but not for the 2021 edition.

In Italy, approximately 56% of university income comes from a block grant provided by the Ministry of University and Research (MUR). Of this grant, 30% is allocated based on the results of a periodic evaluation exercise—the VQR—conducted by the National Agency for the Evaluation of Universities and Research (ANVUR).

The VQR has now entered its fourth cycle, maintaining a consistent evaluation framework with only minor adjustments over time. This structure closely mirrors that of the RAE/REF series, which ANVUR adopted as a benchmark from the outset. Universities approach the process with a clear understanding of its strategic significance and the need to perform at their highest level.

In the most recent edition, VQR 2020–2024, each university was required to submit an average of 2.5 research products per department/school[1] member—ranging from a minimum of 1 to a maximum of 4 per individual—, covering the evaluation period from 2020 to 2024. To maximize their overall score, institutions had to carefully select their strongest outputs. Additionally, they had the unspoken option to reassign faculty across departments before the official cut-off date of November 1, 2024.

In the present work, we want to investigate the possibility for universities to "game" the VQR, particularly by varying the distribution of research staff among its constituent

[1] From now on "department".

departments, at the cut-off date. We will try to model this "game" in order to test whether and to what extent the strategic use of the "degrees of" freedom provided by the VQR can affect the payoff (i.e., the total score) of a university.

To this end, the following section reviews the relevant literature to establish the theoretical framework of the study. In section "The Italian academic system and the national research assessment exercise (VQR)", after a brief overview of the Italian academic system and the VQR, we present our research hypotheses. section "Data and methods" outlines the methodology used to test these hypotheses, while Section "Results" reports the results of the analysis conducted on a sample of Italian universities. Finally, Section "Discussion and conclusions" provides a discussion of the findings and presents our conclusions.

Institutional gaming in response to national research assessment exercises

A growing list of countries have strengthened competitive mechanisms in the research sector, mainly through selective funding and merit-based access to public resources. The introduction of performance-based research funding systems (PBRFS) based on the outcomes of periodic research assessment exercises serves various aims, among which are the stimulation of continuous improvement in research effectiveness and efficiency (Geuna & Piolatto, 2016; Hicks, 2012; Woelert, 2015). According to Zacharewicz et al. (2019), 16 EU countries have adopted some form of PBRFS. The incentive to improve for research organizations comes from the financial leverage embedded in PBRFS (OECD, 2010; Woelert & McKenzie, 2018) and the reputation deriving from ranking among the tops.

However, the adoption of incentive systems for improvement is shaped by a life cycle in which, in addition to the desired effects intended by the policy maker, a number of undesirable effects also occur (Siverbo et al., 2019) up to the emergence of unethical behaviors and misconduct (Graf et al., 2019). In particular, when evaluation systems rely heavily on bibliometric indicators, scholars and institutions may engage in strategic behaviors aimed at inflating measured performance rather than improving underlying research quality. Empirical evidence shows that such dynamics can lead to systematic distortions in citation practices, including forms of "citation gaming" at the country level (Baccini et al., 2019) and the strategic use of self-citations in response to career incentives (Seeber et al., 2019).

In 1997 the anthropologist Marilyn Strathern, commenting on the fact that British universities were increasingly subject to national scrutiny for teaching, research, and administrative competence, claimed: "Audit does more than monitor—it has a life of its own that jeopardizes the life it audits" (Strathern, 1997). In trying to stimulate continuous improvement, evaluation frameworks end up modifying the subject under evaluation. The transformative nature of the evaluation process influences not only the research practices of individual scholars but also the strategic decisions of groups, departments, and research institutions as a whole.

A significant body of literature was developed after the introduction in the UK of the first national research evaluation exercise the (RAE) and its successor, the Research Excellence Framework (REF). The most copious part of that literature examines how these exercises have impacted on individual academics, particularly according to their own perception (Rand, 2021). Although the topic has been frequently mentioned in the literature, there is relatively little empirical research on how universities have responded—particularly

in terms of institutional strategies aimed at 'gaming' the system (Broadbent & Laughlin, 2009; Elton, 2000; Smith et al., 2011).

We must go back to Pfeffer and Salancik (1978) for tracing the theoretical foundations of the external control of organisations through mechanisms of resource dependence. In order to understand how universities respond to external pressure to secure resources (by adopting, among others, gaming strategies), it may also be useful to refer to the principal-agent theory (Jensen & Meckling, 1976). In this case, university research managers (agents) interact with the bodies in charge of evaluation and, hence, funding (principals), and this interaction may lead to opportunistic behaviour on the part of the former in order to maximise the evaluation scores and funding allocated by the latter.

Commissioned by the UK Minister of Universities and Science, in 2015 Lord Nicholas Stern conducted an independent review of the REF (Stern, 2016a, 2016b). Although no empirical evidence is presented, he argued that "Much of the behaviour identified as gaming was linked to the hiring of staff to enhance institutions' REF returns. This included the recruitment of staff on fractional contracts who were based at institutions abroad and played a very limited role in the UK institutions for which they were, submitted; and the movement of staff to new institutions shortly before the REF census date".

The study of the effects of the introduction of the RAE/REF in the UK focused mainly on the observation of recruitment practices. De Fraja et al. (2019), developing a maximization model for academic departments, conclude that 'when hiring or responding to outside offers prior to the REF census date, institutions should value more a researcher with a stellar publication record', given the positive effects this would have on the recruiting university's ranking. Furthermore, the recruitment of a "star" academic if on the one hand ensures the recruiting university the expendability of its scientific portfolio (thanks to the "portability" mechanism), on the other hand also guarantees a productivity improvement of the existing staff (incumbents) especially where collaboration occurs (Agrawal et al., 2017).

Recently, Harris et al. (2025) used faculty panel data for universities in the Russell Group (HESA, 2022) and found that in "the first half of the cycle (the period leading-up to the RAE 2008 and REF 2014 census dates) there was strong evidence that the hiring of professors increased while at the same time the number of professors exiting fell. After the 2008 and 2014 census dates (the second half of the cycle), hiring declined substantially while exiting increased". According to the authors, "these same patterns were not visually evident leading-up to the 2021 REF which suggest that the rule changes made after the 2014 REF, to combat institutional 'gaming', had the desired effect".

In the Italian academic system, the limited flexibility of the academic labour market makes it particularly challenging to adopt strategic recruitment practices within the short timeframes imposed by evaluation cycles. Conversely, numerous studies have highlighted the methodological flaws of such evaluation exercises over the past 15 years and the resulting distortions they may produce (Abramo & D'Angelo, 2015; Abramo et al., 2010; De Nicolao, 2022). The only large-scale empirical analysis on institutional gaming examined the potential incentives introduced by the first VQR, which penalized universities based on the proportion of scientifically inactive faculty members (Abramo et al., 2019). The authors explored the post-evaluation increase in intramural collaborations involving previously unproductive professors, investigating whether highly productive colleagues might have extended gift authorship by allowing their inactive peers to co-sign a small number of their publications. However, the study found no evidence supporting this hypothesis—let alone any indication that such practices were orchestrated by deans or research managers.

In this paper, we aim to further investigate the effects of the Italian VQR, particularly in light of a change introduced in its most recent 2020–2024 edition, which may open the

door to institutional gaming through the strategic reallocation of faculty across departments. Accordingly, the following section offers a concise overview of the Italian academic system and the VQR, setting the stage for the formulation of a series of research hypotheses related to this potential behavior.

The Italian academic system and the national research assessment exercise (VQR)

In Italy, the Ministry of University and Research (MUR) recognizes a total of 99 universities as authorized to confer legally valid academic degrees. Approximately 30 of these are small, private, special-focus institutions, of which 10 operate exclusively through e-learning. Six are Schools for Advanced Studies (*Scuole Superiori*), which are highly selective institutions dedicated to exceptionally talented students, characterized by small faculties and limited enrolment per degree program.

More than 90% of faculty are employed at public universities, which are predominantly funded by the MUR through a block grant known as the *Fondo di finanziamento ordinario* (FFO). In 2023, the FFO amounted to approximately €9 billion, representing about 56% of the total income of the Italian university system. Until 2009, this core government funding was input-oriented, distributed proportionally based on institutional size and disciplinary composition to ensure equitable support across all universities. A significant shift occurred in 2009 following the first national research assessment exercise, when a rewarding FFO share—the *quota premiale*—began to be allocated based on research performance. Today, periodic research assessment exercises (VQRs) inform the distribution of 30% of the FFO.

To support this new performance-based funding model, the National Agency for the Evaluation of Universities and Research (ANVUR) was established in 2010. Among its key responsibilities is the organization of the VQR at regular intervals.

The first VQR covered the period 2004–2010 and introduced a hybrid methodology combining bibliometrics and peer review to evaluate the quality of research produced by state and legally recognized non-state universities, as well as research institutions under the MUR's jurisdiction. Subsequent editions—2011 to 2014 and 2015 to 2019—refined the evaluation criteria, placing increased emphasis on the societal relevance of research through the inclusion of "third mission" activities. The overall methodological framework of the VQR has remained largely consistent over time, with incremental adjustments to accommodate disciplinary differences and evolving research priorities.

The most recent edition, i.e. VQR 2020–2024, aims to assess research activity carried out during the 2020–2024 period, evaluating both universities and their departments. The evaluation is regulated by Ministerial Decree No. 998 (August 1, 2023)[2] and is structured into 17 disciplinary areas, each assigned to a dedicated Assessment Expert Panel (*Gruppo di Esperti della Valutazione*, GEV),[3] as well as an interdisciplinary panel responsible for

[2] https://www.mur.gov.it/it/atti-e-normativa/decreto-ministeriale-n-998-del-01-08-2023.

[3] In the Italian university system all professors are classified in one and only one field (named the scientific disciplinary sector), grouped into disciplinary areas. Details of such scheme is accessible on http://attiministeriali.miur.it/UserFiles/115.htm, last accessed 07/04/2026. Since the GEVs have taken on exactly the same names as the areas, in the following we will use the two terms (areas/GEVs) indistinctly even though the latter (GEV) represents the aggregation of panel experts operating in the relevant area.

evaluating third mission activities (technology transfer, industry collaboration, public engagement, etc.).

The GEVs comprise 655 national experts, primarily from academia. Participating institutions are required to submit research outputs—mainly, though not exclusively, scientific publications, produced during the evaluation period.[4] Each university must submit a total number of products equal to 2.5 times the number of research staff members per department, ranging from a minimum of 1 to a maximum of 4 per individual. Each researcher must be formally affiliated with a specific department, and their output is evaluated accordingly.

Each, submitted product is assigned to a GEV and reviewed by two experts—either internal or external—based on subject-matter expertise. Reviews follow standardized epistemic criteria: (i) originality, (ii) methodological rigor, and (iii) impact. The product is then assigned a quality class and corresponding score:

(a) Outstanding → 1.0
(b) Excellent → 0.8
(c) Standard → 0.5
(d) Acceptable → 0.2
(e) Not acceptable → 0.

In Social Sciences and Humanities (SSH), evaluation relies solely on peer review. In the Sciences, peer review is informed by bibliometric indicators—citation counts and journal percentile rankings—sourced from Scopus and/or Web of Science (WoS).

The final quality profile of a university is calculated as the average score of its, submitted research products. The portion of the rewarding FFO share received depends on this average score and the number of products, submitted. As such, institutions are incentivized to strategically select products that maximize their overall score.

However, this optimization is subject to several constraints. Each department must submit, for every faculty member, between one and four products, with an average of 2.5 per person. Co-authored works can only be, submitted once per university, with one exception: if the product has fewer than six total authors, two of whom belong to different Departments/Schools, it may be, submitted by both.

The evaluation will be based on the research staff composition as of November 1, 2024, the official cut-off date. ANVUR requires each university to indicate whether internal units to be evaluated will be Departments or Schools and to validate the list of eligible research staff accordingly (ANVUR, 2024).

The authors of this study hypothesize that universities might exploit these "degrees of freedom" to enhance their performance by strategically modifying internal structures, particularly departmental affiliations. This leads to the formulation of the following hypotheses:

H1 *The lower the average research quality of a university's faculty, the greater the incentive to engage in strategic departmental restructuring.*

[4] As said, the VQR also includes the submission of a number of case studies related to third mission activities. In the following, we will ignore this and focus only on research products and, therefore, on the evaluation of research activities only.

H2 *Universities with departments exhibiting higher levels of intra-departmental collaboration have greater potential for improvement via product duplication.*

H3 *The ability to benefit from intra-departmental collaborations is further amplified when such collaborations yield higher-quality research outputs.*

The following sections aim to model this strategic behavior and empirically test these hypotheses.

Data and methods

The mathematical formulation

In this section, we introduce a MILP (mixed-integer linear programming) model designed to maximize the total score of Departments/Schools by varying their composition while following the constraints outlined in the previous section.

The model allows testing different scenarios, including the current scenario and those that permit a specific percentage of researcher movement between departments. The magnitude of the increase in the total score between scenarios is intended to serve as a proxy for the university's incentive for institutional gaming, i.e., engaging in strategic "departmental affiliation changes".

A key modeling issue concerns the representation of research quality. The VQR assigns each product to a discrete quality class, while bibliometric indicators provide a continuous measure of impact.

To account for this, we associate each product p with:

- a continuous score s_{pr},
- a discrete score c_{pr}, obtained by mapping the percentile into VQR-like quality classes.

The model can be solved under two alternative objective configurations:

1. a continuous formulation, where the total continous score is maximized;
2. a lexicographic formulation, where the total discrete score is maximized first, and the continuous score is maximized among optimal solutions.

This allows us to interpret the results both as an upper-bound estimate (continuous case) and as consistent with the VQR evaluation mechanism (lexicographic case).

Table 1 lists the notation used in the mathematical model.

Let z denote the objective function value. Its definition depends on the selected scoring formulation. In particular, we define:

$$z_1 = \sum_{(p,r)\in A} \sum_{d\in D} c_{pr} x_{prd}$$

$$z_2 = \sum_{(p,r)\in A} \sum_{d\in D} s_{pr} x_{prd}$$

Table 1 Model sets, parameters and variables

Sets	D	Departments
	R	Researchers
	P	Products
	A	Pairs (p,r), where $p \in$ P is a product and r $\in$ R is a researcher associated with that product
	B	Pairs (r,d), where $r \in R$ is a researcher, and $d \in D$ is the initial Department to which the researcher is affiliated. This set represents the baseline scenario before any changes to departmental affiliations are made
Parameters	c_{pr}	Quality class score assigned of product p of researcher r
	s_{pr}	Continuous score assigned of product p of researcher r
	t_u	Total number of researchers of university
	n_p	Number of authors of product p
	α	Number of researchers that can be assigned to a different Department
	β	Threshold for constraints definition related to the number of authors per product
	γ	Minimum number of researchers in a Department
Variables	x_{prd}	Binary decision variable equals to 1 if the product p is selected for the researcher r of department d, 0 otherwise
	z_{rd}	Binary decision variable equals to 1 if the researcher r is assigned to department d, 0 otherwise

where z_1 represents the total discrete VQR-like score and z_2 the total continuous score.

The objective function z is defined as follows:

- Continuous formulation: $z = z_2$
- Lexicographic formulation: first maximize z_1, and then maximize z_2 among optimal solutions.

The model is therefore written in compact form using a generic objective z, while its specific definition depends on the chosen scoring configuration.

In the lexicographic formulation, this is implemented through a two-step optimization procedure. First, the model maximizes z_1. Then, the model maximizes z_2 subject to all original constraints and the additional constraint:

$$\sum_{(p,r)\in A} \sum_{d\in D} c_{pr} x_{prd} \geq z_1^*$$

where z_1^* is the optimal value obtained in the first step. Considering the above notation, the problem can be formulated as follows:

$$max z \tag{1}$$

$$\sum_{p\in P:(p,r)\in A} x_{prd} \geq z_{dr} \qquad \forall d \in D, r \in R \tag{2}$$

$$\sum_{p\in P:(p,r)\in A} x_{prd} \leq 4 z_{dr} \qquad \forall d \in D, r \in R \tag{3}$$

$$\sum_{p\in P:(p,r)\in A} x_{prd} \leq 2.5 \sum_{i\in R} z_{dr} \qquad \forall d \in D \tag{4}$$

$$\sum_{r\in R} z_{dr} \geq \gamma \qquad \forall d \in D \tag{5}$$

$$\sum_{d\in D} z_{rd} = 1 \qquad \forall r \in R \tag{6}$$

$$\sum_{r\in R:(p,r)\in A} x_{prd} \leq 1 \qquad \forall r \in R \tag{7}$$

$$\sum_{d\in D} \sum_{r\in R:(p,r)\in \mathrm{A}} x_{prd} \leq 2 \qquad \forall p \in P : n_p < \beta \tag{8}$$

$$\sum_{d\in D} \sum_{r\in R:(p,r)\in \mathrm{A}} x_{prd} \leq 1 \qquad \forall p \in P : n_p \geq \beta \tag{9}$$

$$\sum_{d \in D} \sum_{r \in R:(r,d) \in B} z_{rd} \geq t_u - \alpha \tag{10}$$

The objective function (1) maximizes the total score of Departments according to the selected formulation. Constraints (2), (3) and (4) guarantee that each department submits at least one and at most four products per Researcher, with a total limit of 2.5 times the number of researchers in the Department/School, respectively.

Note that the VQR establishes a target submission volume (2.5 products per researcher on average), but effective submission is bounded by the availability of eligible outputs. The model, therefore, encodes this as an upper bound constraint. Missing products reflect the structural scarcity of assignable outputs and imply a lower total score, consistent with the evaluation mechanism.

Constraints (5) and (6) ensure that each department must formally assign at least γ researchers, and each researcher must belong to a specific department. Constraints (7), (8) and (9) regulate the submission of co-authored products, allowing each product to be, submitted only once per university (and thus per department), except for products with fewer than $\beta = 6$ authors, two of whom belong to different departments. In such cases, the product can be, submitted by up to two departments. Finally, Constraint (10) regulates the movement of researchers between departments, ensuring that such movements remain within allowable limits.

The numerical resolution of this problem is achieved using an integer linear programming model implemented in Python and solved by Gurobi.

All Python scripts implementing the MILP model and simulation procedures are publicly available at https://github.com/Alecxio89/InstitutionalGaming_VQR. Due to licensing restrictions associated with the Web of Science data, the raw bibliometric dataset cannot be shared; however, synthetic data reproducing the model structure is provided to ensure computational reproducibility.

Data and sources

In order to test our research hypothesis, we applied the above model, under the two objective configurations previously defined, to a sample of national universities. Data on their faculty were extracted from the database on Italian university personnel, maintained by the MUR.[5] We considered as cut-off date the 31/12/2023.

The bibliometric dataset is extracted from the Italian Observatory of Public Research, a database developed and maintained by the present authors and derived under license from the Web of Science of Clarivate Analytics. Beginning from the raw data of the WoS, and applying a complex algorithm to reconcile the author's affiliation and disambiguation of the true identity of the authors, each publication (article, article review and conference proceeding) is attributed to the university scientist or scientists that produced it (D'Angelo et al., 2011). For the purpose of the proposed analyses, the publications considered are those indexed in WoS in the 2018–2022 period. This time frame provides an adequate citation window for assessing the impact of publications. At this purpose, each publication is assigned a 0–100 score as the citation percentile derived considering all publications in the same year and subject category.

[5] https://cercauniversita.mur.gov.it/php5/docenti/cerca.php

Table 2 Dataset of the analysis

University code	Size	No. of Depts	No. of authors	No. of WoS publications	Avg publications per author
VL	Very large	31	2545	51,804	20.4
L	Large	18	1144	21,899	19.1
M1	Medium	8	585	8144	13.9
M2	Medium	9	443	8841	20.0
M3	Medium	9	532	13,670	25.7
S1	Small	2	162	9187	56.7
S2	Small	3	143	1946	13.6

It is important to clarify that our model does not attempt to replicate the informed peer-review process of the VQR in its entirety. Rather, citation percentiles are used as a continuous bibliometric proxy for research quality in scientific fields, consistent with the bibliometric support explicitly incorporated in the VQR evaluation framework. Accordingly, the model should be interpreted as a structural vulnerability analysis under a bibliometric surrogate of quality.[6]

Citations are observed at the cut-off date (31/12/2023). A publication with a score of 75, has received (at that date) a number of citations that places it on the threshold of the first quartile of the distribution. In essence, only 25 % of the publications in the same year and subject category received a number of citations greater than or equal to those received by the publication under consideration.

At the same time, the time window was chosen to realistically simulate outcomes for a five year VQR exercise (as the latest 2020–2024 edition), as gaming incentives should not be impacted by the publication window.

Our dataset consists of 7 universities of different sizes. As shown in Table 2, the largest has 31 departments and more than 2500 researchers producing, on average, 20 publications each in the 2018–2022 period. Next, we find a university with 18 departments and 1144 researchers. Then, 3 medium-sized universities with 8–9 departments and varying research staff between 443 and 585 units. Finally, two small universities with 2–3 departments and a research staff of about 150 units. The first one (S1) registered 56.7 publications per author (the highest value among the 7 universities), and the second only 13.6 (the lowest value). This does not necessarily imply a difference in the "productivity" of their researchers; it is simply possible that the research staff of the former is engaged in more publication-intensive fields than the latter.

Note that the seven universities were selected through a purposive stratified sampling strategy aimed at ensuring heterogeneity in institutional size and organizational structure (very large, large, medium, and small institutions). The objective was not statistical representativeness of the Italian university system, but rather to test the structural behavior of the optimization model across diverse configurations. Accordingly, the analysis should be

[6] To assess robustness, we also conducted simulations using a discretized scoring regime that maps citation percentiles into five classes aligned with the VQR scoring system (1.0; 0.8; 0.5; 0.2; 0).

Table 3 Results of simulations for the AS-IS configuration of the seven universities in the dataset

University code	Products to submit	Products selected	Missing products	Total score	Avg score
VL	6354	6053	301 (4.7%)	426,196.6	70.4
L	2856	2582	274 (9.6%)	190,124.0	73.6
M1	1462	1280	182 (12.4%)	86,588.5	67.6
M2	1104	1013	91 (8.2%)	78,824.1	77.8
M3	1330	1328	2 (0.2%)	120,867.4	91.0
S1	405	405	0 (0.0%)	37,275.3	92.0
S2	357	287	70 (19.6%)	19,732.7	68.8

interpreted as a vulnerability and mechanism exploration exercise rather than a statistically generalizable estimate of gaming magnitude at the national level.

Results

This section presents the results of the optimizations applied to seven Italian universities. The aim was to evaluate the potential for institutional 'gaming' through internal reallocation of research staff among departments, considering the rules of the VQR 2020–2024. First, the baseline AS-IS configuration—without any strategic changes—was considered, and then it was compared to alternative scenarios allowing limited staff movements.

Table 3 shows the outcomes of the MILP optimization under the AS-IS configuration, i.e., without any strategic rearrangement of departmental affiliations. For each university, the number of required products was determined according to the VQR rule of submitting, on average, at most 2.5 products per researcher. The *Missing products* column indicates the gap between the maximum number of eligible products that could be selected under current departmental assignments and the number actually selected.[7] This difference does not result from an optimization failure, but rather a consequence of structural constraints: in several departments, the pool of assignable products fell below the quota because some researchers had not produced a sufficient number of outputs during the evaluation period, or because existing products were already assigned to other departments under co-authorship rules. These performance gaps have clear implications for institutional gaming strategies. Universities with large deficits of eligible products (e.g., S2, M1, L) may benefit substantially from reconfiguring departmental affiliations to enable better product selection or exploit VQR duplication rules.

Before presenting the results of the gaming scenarios, we briefly assess the robustness of the model with respect to the scoring formulation. In particular, we compare the continuous formulation, with the lexicographic formulation, which prioritizes VQR-like discrete classes and then refines the solution using the continuous score.

The comparison shows that the two formulations yield nearly identical solutions across all instances. Differences in total score are consistently below 1%, with only one scenario

[7] Missing products do not reflect the infeasibility of the optimization model, but rather insufficient eligible outputs under departmental and co-authorship constraints. Since funding depends on both average score and number of products submitted, such under-submission effectively penalizes institutions.

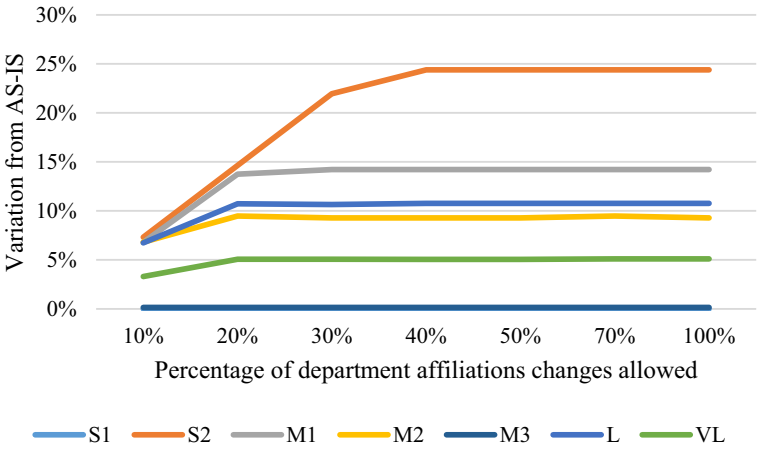


Fig. 1 Increase of the number of products selected, resulting from changes in department affiliations, for the seven universities in the dataset

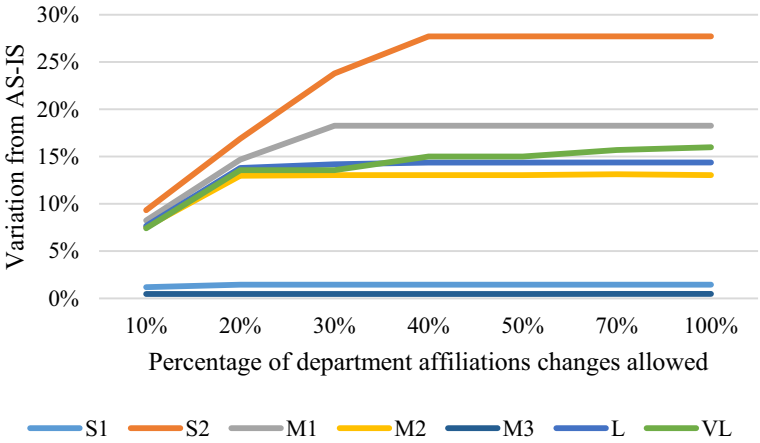


Fig. 2 Increase of total score, resulting from changes in department affiliations, for the seven universities in the dataset

exhibiting a slightly larger deviation (approximately 1%), where the continuous formulation yields a marginally higher score.

This indicates that the optimization outcomes are not driven by the choice of scoring representation but rather reflect structural properties of the problem and the evaluation constraints. Therefore, in what follows, we report results based on the continuous formulation, which provides a more informative representation.

Figures 1 and 2 illustrate the variation in the number of selected products and the corresponding changes in the total score resulting from institutional "gaming". These results suggests that universities with lower initial average scores tend to achieve greater gains in both the number of selected products and overall score. This confirms Hypothesis H1—that universities starting from a weaker position have more to gain from internal reorganization. Notably, S2, with the lowest AS-IS average score (68.8) and the highest percentage

Table 4 Indicators of product duplicability, quality, and completeness for the seven universities in the dataset

University code	% Duplicable products duplicabili	% High-quality duplicable	Missing products	Avg score
VL	9.0	39.0	301 (4.7%)	70.4
L	20.8	25.6	274 (9.6%)	73.6
M1	36.0	29.7	182 (12.4%)	67.6
M2	27.8	21.5	91 (8.2%)	77.8
M3	1.4	5.3	2 (0.2%)	91.0
S1	56.8	7.0	0 (0.0%)	92.0
S2	29.4	34.3	70 (19.6%)	68.8

of missing products (19.6%), achieves the most substantial improvement in the simulations. Conversely, S1 and M3, with very high initial scores and nearly complete submissions, show negligible variation in both indicators.

Table 4 provides insight into the internal collaboration dynamics that influenced the observed improvements. It reports, for each university, the percentage of *duplicable products* (i.e., those eligible for submission by two departments according to VQR rules), and the percentage of those with a score higher than the score obtained by the university in the AS-IS scenario.

A product is defined as potentially *duplicable* under reassignment if it has between 2 and 5 total authors and includes at least two researchers affiliated with the same department in the AS-IS configuration. Such products may become eligible for dual submission under VQR rules if departmental reassignment creates cross-department co-authorship configurations. The indicator, therefore, captures structural convertibility potential rather than direct rule-based duplicability in the baseline configuration.

These results support Hypothesis H2: universities with higher rates of intra-departmental co-authorship (proxied by duplicable products) show greater scope for performance improvement. For example, M1 and S2 have both high duplicability and large score gains in simulations. However, duplicability alone is insufficient—as shown by S1, whose products are largely duplicable (56.8%) but of relatively low quality (only 7.0% above the median), resulting in negligible improvement. This supports Hypothesis H3: meaningful gains from internal gaming require not only the opportunity to duplicate outputs, but also that those outputs are of high scientific quality. A further example is University VL, which has relatively low duplicability (9.0%) but a high percentage of high-quality duplicable products (39.0%). Despite fewer opportunities for duplication, it still shows substantial improvement, confirming that the quality of co-authored products is a stronger predictor of gain than quantity alone.

To illustrate the practical impact of institutional gaming, Table 5 presents the results of two simulated scenarios applied to University M3: the first allows for a 10% reallocation of researchers across departments (Game 1), while the second simulates the aggregation of all departments into a single school (Game 2). Although M3 already performs at a high level in the AS-IS scenario and has limited scope for improvement, both configurations result in measurable—though modest—gains.

Table 5 Results of simulations and differences between the AS-IS and two gaming scenarios for M3 University

Scenario	Description	Total score	Avg score
AS-IS	–	120,867	91.0
Game1	10% of department affiliations changes allowed	121,430 (+0.47%)	91.3 (+0.31%)
Game2	Aggregation of departments into the School	121,336 (+0.39%)	91.2(+0.24%)

Discussion and conclusions

This study investigated the incentives and potential for institutional 'gaming' among Italian universities participating in the VQR 2020–2024 national research evaluation exercise. By applying a mixed-integer linear programming model to bibliometric data and staffing configurations across seven Italian universities, our findings confirm that strategic internal reorganization can significantly influence evaluation outcomes.

The study adopts a positive analytical perspective grounded in resource dependence and principal–agent theory. Universities are modeled as rational actors responding to the incentive structure embedded in the evaluation framework. From this perspective, strategic reallocation of faculty is not inherently unethical; rather, it represents a predictable behavioral response to institutional incentives. The normative question of whether such behavior should be constrained is ultimately a policy-design issue. Our contribution is therefore diagnostic rather than prescriptive: we identify structural vulnerabilities in the evaluation architecture that may generate unintended optimization behaviors.

The outcomes of our analysis substantiate all three proposed hypotheses. Firstly, institutions with lower baseline research performance, reflected in their average scores, exhibit stronger incentives to restructure internally (Hypothesis H1). Particularly, universities facing substantial deficits in eligible research outputs demonstrate pronounced gains through strategic faculty reassignment. These results highlight how lower-performing universities may resort to restructuring as a means to optimize their evaluation scores strategically.

Secondly, universities characterized by higher intra-departmental collaboration possess greater opportunities to exploit VQR rules regarding duplicable research outputs (Hypothesis H2). However, our findings importantly nuance this point: high levels of collaboration alone do not necessarily yield significant evaluation gains. It is the quality of these collaborative outputs that ultimately drives substantial improvements in the universities' evaluation scores, supporting Hypothesis H3. High-quality co-authored publications emerged as a pivotal factor, reinforcing the crucial role of scientific excellence over mere numerical productivity.

Further analysis revealed that even institutions with relatively fewer opportunities for duplication could achieve significant improvements when their duplicable products were of high scientific quality. For example, universities with moderate duplication potential but high-quality collaborative outputs demonstrated notable gains in evaluation scores. This finding emphasizes the strategic importance of quality-driven collaborations, not only in maximizing evaluation scores but also in genuinely enhancing institutional research capabilities.

Additionally, the model simulations illustrated practical scenarios—such as allowing modest departmental reassignment (10%) or aggregating multiple departments into fewer, larger units—that yielded measurable performance improvements even in high-performing universities. Although these gains were modest, they underline the pervasive nature of incentives created by the evaluation framework, affecting not only institutions underperforming initially but also those already performing at high standards.

Another relevant insight concerns the role of the scoring structure in shaping institutional incentives. Although the model was implemented under both a continuous formulation and a VQR-consistent lexicographic formulation, the resulting solutions are nearly identical.

This suggests that the observed incentives for institutional gaming are not driven by the specific representation of research quality (continuous vs. discrete), but rather by the combinatorial structure of the evaluation constraints. In this sense, the identified vulnerabilities should be interpreted as structural features of the evaluation framework and are therefore likely to persist under alternative scoring specifications.

Several limitations must be considered when interpreting the results. First, our analysis was constrained by the dataset's limited scope—only seven universities—potentially restricting broader generalizations. The variability among these institutions concerning size, departmental structures, and disciplinary coverage implies caution when extrapolating results to the entire Italian university system. To this end, we adopted a targeted stratified sampling strategy designed to ensure diversity in terms of size and organisational structure, since the aim was not to achieve statistical representativeness of the Italian university system, but rather to examine the structural behaviour of the optimisation model under different configurations. On the other hand, it should be noted that, for the purposes of the study, we could also have analysed synthetic plausible, but not real, datasets: the validity of the results and the conclusions drawn from them would not have changed.

Additionally, the model assumes rational optimization behavior without incorporating possible institutional constraints such as internal politics, resistance to restructuring, or the administrative feasibility of large-scale reassignment of faculty members. These real-world barriers could significantly limit the extent to which universities could realistically 'game' the system.

Furthermore, our reliance on bibliometric data from the Web of Science potentially introduces discipline-specific biases, underrepresenting the full spectrum of research outputs, particularly in the Humanities and Social Sciences. Therefore, the potential for gaming could differ significantly in institutions with varied disciplinary profiles.

Finally, we recall that in our simulation the time framework differs from that of the 2020–2024 VQR edition, although we do not expect this to affect the results of the analysis. More generally, the whole of our empirical framework does not attempt to replicate the informed peer-review process of the VQR in its entirety. Citation percentiles (and the five VQR-like classes associated to percentile thresholds) are used as a continuous bibliometric proxy for research quality, consistent with the bibliometric support explicitly incorporated in the VQR evaluation framework. Accordingly, the model should be interpreted as a structural vulnerability analysis under a bibliometric surrogate of quality.

Given these limitations, caution is advised regarding the generalization of these results beyond the sample universities and beyond Italy. Although the VQR closely mirrors other international frameworks like the UK's REF, differences in national academic cultures, institutional governance structures, and labor market flexibility imply that results could differ substantially across contexts. Thus, further comparative international research would be valuable to better understand the universality or specificity of these findings.

Our study underscores critical implications for policymakers and evaluation agencies such as ANVUR. Awareness of the potential for institutional gaming should prompt careful reconsideration of evaluation criteria and submission rules to mitigate strategic behaviors that distort genuine research excellence assessments. Ensuring that evaluation frameworks reward authentic research productivity and collaboration rather than administrative maneuvering will preserve their credibility and effectiveness.

Moreover, our results suggest a pressing need for transparency and accountability in departmental assignment and evaluation processes, possibly through the introduction of stricter controls or penalties against clear gaming behaviors. An ongoing dialogue among policymakers, universities, and academic communities about these vulnerabilities is crucial.

Future research should address several open questions. First, expanding this analysis to include a larger, representative sample of universities would improve the robustness and generalizability of the findings. Comparative analyses across different countries with similar evaluation exercises could also elucidate the factors that promote or inhibit institutional gaming behaviors universally.

Further qualitative research into university decision-making processes could provide deeper insights into the dynamics underpinning strategic restructuring decisions, highlighting the human and organizational complexities involved. Lastly, exploring alternative bibliometric and qualitative indicators could enhance the evaluation framework, reducing its susceptibility to manipulation and enhancing fairness and transparency in performance-based funding systems.

Funding Open access funding provided by Università degli Studi di Roma Tor Vergata within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

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., & Di Costa, F. (2019). When research assessment exercises leave room for opportunistic behavior by the subjects under evaluation. *Journal of Informetrics, 13*(3), 830–840. https://doi.org/10.1016/j.joi.2019.07.006

Abramo, G., D'Angelo, C. A., & Viel, F. (2010). Peer review research assessment: A sensitivity analysis of performance rankings to the share of research product evaluated. *Scientometrics, 85*(3), 705–720. https://doi.org/10.1007/s11192-010-0238-0

Agrawal, A., McHale, J., & Oettl, A. (2017). How stars matter: Recruiting and peer effects in evolutionary biology. *Research Policy, 46*, 853–867.

ANVUR (2024). *BANDO Valutazione della Qualità della Ricerca 2020–2024.* https://www.anvur.it/wp-content/uploads/2023/10/Bando-VQR-2020-2024_31ottobre.pdf.

Baccini, A., De Nicolao, G., & Petrovich, E. (2019). Citation gaming induced by bibliometric evaluation: A country-level comparative analysis. *PLoS ONE, 14*(9), Article e0221212. https://doi.org/10.1371/journal.pone.0221212

Broadbent, J., & Laughlin, R. (2009). Performance management systems: A conceptual model. *Management Accounting Research, 20*(4), 283–295.
D'Angelo, C. A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. *Journal of the American Society for Information Science and Technology, 62*(2), 257–269. https://doi.org/10.1002/asi.21460
De Fraja, G., Facchini, G., & Gathergood, J. (2019). Academic salaries and public evaluation of university research: Evidence from the UK Research Excellence Framework. *Economic Policy, 34*, 523–583.
De Nicolao, G. (2022). State-fostered Immaturity? Kant, Galileo, and the Grand Evaluator. In I. De Gennaro, H. Hofmeister, & R. Lüfter (Eds.), *Academic Freedom in the European Context: Legal, Philosophical and Institutional Perspectives* (pp. 333–349). Cham: Springer International Publishing.
Elton, L. (2000). The UK research assessment exercise: Unintended consequences. *Higher Education Quarterly, 54*, 274–283.
Geuna, A., & Piolatto, M. (2016). Research assessment in the UK and Italy: Costly and difficult, but probably worth it (at least for a while). *Research Policy, 45*(1), 260–271.
Graf, L., Wendler, W. S., Stumpf-Wollersheim, J., & Welpe, I. M. (2019). Wanting more, getting less: Gaming performance measurement as a form of deviant workplace behavior. *Journal of Business Ethics, 157*, 753–773. https://doi.org/10.1007/s10551-017-3688-y
Groen-Xu, M., Bos, G., Teixeira, P. A., Voigt, T., & Knapp, R. (2023). Short-term incentives of research evaluations: Evidence from the UK Research Excellence Framework. *Research Policy, 52*(104729), 1–22.
Harris, R., Mate-Sanchez-Val, M., & Ruiz Marín, M. (2025). Institutional 'gaming' involving staff turnover during recent research evaluation exercises by UK Russell Group universities. *Oxford Economic Papers*. https://doi.org/10.1093/oep/gpae032
HESA. (2022). Higher education staff data. Retrieved April 07,2026 from, https://www.hesa.ac.uk/data-and-analysis/staff
Hicks, D. (2012). Performance-based university research funding systems. *Research Policy, 41*(2), 251–261.
Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs, and ownership structure. *Journal of Financial Economics, 3*(4), 305–360.
Marginson, S. (2006). Dynamics of national and global competition in higher education. *Higher Education, 52*(1), 1–39.
OECD, Organisation for economic Co-operation and development. (2010). *Performance-based Funding for Public Research in Tertiary Education Institutions*. OECD Publishing.
Pfeffer, J., & Salancik, G. R. (1978). *The external control of organizations: a resource dependence perspective*. Harper & Row.
Rand (2021). Understanding perceptions of the research excellence framework among UK researchers. Retrieved April 07,2026 from, https://www.rand.org/pubs/research_reports/RRA1278-1.html.
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*(2), 478–491. https://doi.org/10.1016/j.respol.2017.12.004
Siverbo, S., Cäker, M., & Åkesson, J. (2019). Conceptualizing dysfunctional consequences of performance measurement in the public sector. *Public Management Review, 21*, 1801–1823.
Smith, S., Ward, V., & House, A. (2011). 'Impact' in the proposals for the UK's Research Excellence Framework: Shifting the boundaries of academic autonomy. *Research Policy, 40*, 1369–1379.
Stern, N. (2016b). *Research Excellence Framework Review: Evidence Report*. Retrieved April 07,2026 from https://www.gov.uk/government/publications/research-excellence-framework-review
Stern, N. (2016a). *Building on Success and Learning from Experience: An Independent Review of the Research Excellence Framework*. Retrieved April 07,2026 from, https://www.gov.uk/government/publications/research-excellence-framework-review
Strathern, M. (1997). 'Improving ratings': Audit in the British university system. *European Review, 5*(3), 305–321.
Woelert, P. (2015). Governing knowledge: The formalization dilemma in the governance of the public sciences. *Minerva, 53*(1), 1–19.
Woelert, P., & McKenzie, L. (2018). Follow the money? How Australian universities replicate national performance-based funding mechanisms. *Research Evaluation, 27*(3), 184–195.
Zacharewicz, T., Lepori, B., Reale, E., & Jonkers, K. (2019). Performance-based research funding in EU member states-a comparative assessment. *Science and Public Policy, 46*(1), 105–115.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.