Scientometrics
https://doi.org/10.1007/s11192-025-05240-2

Ranking Departments based on research quality: a statistical evaluation of the ISPD indicator

Federica Galli[1] · **Fedele Greco**[1]

Received: 6 February 2024 / Accepted: 10 January 2025

Abstract
Given the relevance of performance-based funding for the Italian University System, in this paper, we analyse the statistical properties of the ISPD indicator used to provide a ranking of the best 180 Italian Departments. Besides the adoption of a standardization procedure and subsequent aggregation, the main flow of the ISPD indicator appears to be its tendency to polarization, which has important consequences on the ability of such indicator to effectively discriminate between outstanding Departments as opposed to those just above the average. Implications on funding allocation are discussed using data on the 2011-2014 Italian research assessment exercise and some proposals on alternative indicators that solve some of the critical aspects of ISPD are sketched.

Keywords Departments of excellence · Italian research assessment · Polarization

1 Introduction

Italy funds its Universities predominantly through state contributions provided by the Ordinary Financing Fund (Fondo per il Finanziamento Ordinario delle Universitá, FFO). A portion of the FFO, representing between 7% and 30% of the total allocation, is distributed based on performance-based criteria. Of this performance-based funding, 60% is determined by research quality, as evaluated through the national research assessment exercise, the Valutazione della Qualitá della Ricerca (VQR). Besides the FFO performance-based share, a source of financial support to Italian Universities is assigned to Departments that stand out for the quality of the research produced, called Departments of Excellence. In order to select the Departments of Excellence, a preliminary raking of the best 350 Italian Departments is defined on the basis of an indicator computed starting from the VQR results named Standardised Indicator of Departmental Performance (Indicatore Standardizzato della Performances Dipartimentale, ISPD) conceived by Giacomo Poggi (Poggi & Nappi, 2014; Poggi, 2015). The final ranking of the 180 Departments of Excellence is obtained by combining the ISPD score and the evaluation of a departmental development

✉ Federica Galli
federica.galli14@unibo.it

[1] Department of Statistical Sciences “Paolo Fortunati”, University of Bologna, Via delle Belle Arti 41, Bologna, Italy

project submitted by the 350 eligible Departments. In the period 2018-2022, the 180 Departments selected have received, depending on their size, an additional financial contribution between 1.620 and 1.080 million euros per year for five years (Law 11 December 2016, n.232).

The first Italian research assessment experiment dates back to VTR (Valutazione Triennale della Ricerca, or Three-year Research Assessment) performed in 2000-2003. However, due to the unclear procedure for the selection of products and the lack of impact on funding, the influence of the VTR was very limited. The first large-scale research assessment was performed in 2004-2010 and evaluted by the National Agency for the Evaluation of Universities and Research Institutes (ANVUR), first created in 2006 becoming operational in 2010. Panels of evaluation experts were created for the 14 different Research Areas[1] defined as aggregations of disciplines called Scientific Disciplinary Sectors (SDS) and evaluations were mainly based on bibliometric criteria for STEM disciplines[2] (Areas 1-9) and peer review for Areas 10-14, with some exceptions. Products were assigned to different classes of merit corresponding to scores from 0 to 1 and final indicators were computed based on different aggregations by SDS, Department and University. The subsequent research assessments took place in the years 2011-2014 and 2015-2019 incorporating some improvements and modifications learning from previous exercises (for a detailed description of the different VQR exercises see Bonaccorsi (2020)). Starting in 2017, the VQR results were also used to assign the Departments of Excellence related funds other than allocating the FFO performance-based share.

As anticipated before, starting from the scores assigned to each product evaluated in the VQR, the best Italian Departments are identified based on the ISPD indicator, a standardized indicator of departmental performance. As underlined by the Italian National University Council (CUN) in the document n.20400 of 18/7/2017, the use of an ad-hoc indicator to produce a ranking of Italian Departments is needed due to the heterogeneity of their composition in terms of SDS and in the evaluation of publications within the various sectors. Thus, the ISPD relies on a standardization procedure of the scores obtained by Departments aimed at taking into account the diversity in the average scores and the dispersion of the results in the different SDSs. Once obtained the standardized score for each Department, the ISPD is based on the probability that a virtual Department with the same member composition as the d-th Department will receive a worse evaluation by permuting the affiliates of the SDSs that make up the d-th Department in all Italian public institutions. However, as highlighted by the note published by CUN, two Departments with standardized scores very close to each other can instead have extremely different ISPD values.

Several works have raised strong criticisms of the methods and indicators used in the VQR exercises. In particular, authors concentrated on analysing the operational procedures followed to select and grade products (Abramo et al., 2014; Abramo & D'Angelo, 2015, 2016; Baccini & De Nicolao, 2016; Demetrescu et al., 2020; Marzolla, 2015; Baccini & De Nicolao, 2022; Demetrescu et al., 2020) detecting different issues such as the

[1] The Research Areas are defined as follows: Area 1 (Mathematics and Computer Sciences); Area 2 (Physics); Area 3 (Chemistry); Area 4 (Earth Sciences); Area 5 (Biology); Area 6 (Medicine); Area 7 (Agricultural and Veterinary Sciences); Area 8 (Architecture and Civil Engineering); Area 9 (Industrial and Information Engineering); Area 10 (Ancient History, Philology, Literature and Art History); Area 11 (History, Philosophy, Pedagogy, Psychology); Area 12 (Law); Area 13 (Economics and Statistics); Area 14 (Political and Social Sciences). Each area is then divided into more specific scientific sectors called SDS.

[2] Peer review was also used in STEM disciplines for articles that were either too recent to have received citations or had a too short citation window to ensure reliable results.

low number of papers evaluated via peer review, the way in which coauthorships are handled, the incorrect and anachronistic use of journal metrics and the misleading normalization and composition methods used to combine different bibliometric indicators in order to assign each product to a class of merit. Recently, researchers have begun evaluating the impact of research assessment on academic behavior, highlighting potential adaptations to previous evaluation outcomes and distortive effects, such as strategic journal targeting and the selection of topics driven by utilitarian ends rather than by intellectual curiosity (Spanò et al., 2024; Akbaritabar et al., 2021; Checchi et al., 2020). However, to our knowledge, there are yet no works analysing the ISPD indicator used by ANVUR to rank Departments as well as aggregate indicators used to distribute the FFO-performance based funds.

Given the strong relevance of the ranking produced by the ISPD in allocating additional funds to Italian Universities, which still rely on limited public financial support compared to other European countries,[3] in this paper, we analyze the statistical properties of this indicator, we discuss the main implications resulting from the procedures adopted to define it such as permutation, standardization and dependence on the size of the collective, and we compare it with some alternative indicators. After discussing the statistical properties of ISPD, we focus on their implications on funding, showing that the project score plays a crucial role in the final ranking of the Departments of Excellence: this is mainly due to the tendency of ISPD to deliver very similar scores for above-the-average Departments. We demonstrate that such scores show remarkably lower heterogeneity with respect to those obtained by several other indicators.

Statistical analyses are based on SDS and Department-level data provided in the tables attached to the Research Area reports from the VQR 2011-2014. A more up-to-date analysis should have involved data from the latest VQR exercise: unfortunately, reports from VQR 2015-2019 do not provide disaggregated data at SDS and Department level, with a striking loss of transparency and reproducibility of the results with respect to VQR 2011-2014. Nonetheless, data from VQR 2011-2014 are perfectly suitable for studying the *statistical properties* of the ISPD indicator: as a matter of fact, the technical details concerning the ISPD are unchanged between the two evaluation exercises as witnessed by the VQR 2011-2014 (ANVUR, 2017b) and VQR 2015-2019 (ANVUR, 2022b) methodological notes, which are identical. The two assessment exercises only differ with respect to the classification of the research products (A-F in VQR 2011-2014, A-E in VQR 2015-2019) and with respect to the scores associated with each category. Moreover, affiliates were asked to present 2 scientific products for VQR 2011-2014 and 3 scientific products for VQR 2015-2019. It is worth noting that the data used in this paper do not allow to exactly reproduce the ISPDs values officially published in 2017 when selecting the Departments of Excellence because of shifts in the composition of the Departments in time and because of the non-exhaustiveness of the data provided by the tables available online due to confidentiality restrictions, since data disaggregated by SDS and Department where not published for SDSs with less than 5 scientific products in a given Department.

In the remainder of the paper, Sect. 2 describes the data used for the analysis and provides some descriptive statistics. The statistical properties of the ISPD indicator are discussed in Sect. 3, while Sect. 4 provides an empirical application using data from the 2011-2014 Italian research assessment. In Sect. 5, some concluding remarks are sketched.

[3] Although the scope of this paper is limited to Italy, a brief comparison of performance-based funding systems across EU countries is included in Appendix A.

2 Data

The data referring to the VQR 2011-2014 have been retrieved from the ANVUR website. Information is provided by Area differentiating by Department and SDS. All data tables and R codes useful to reproduce the results of this analysis can be downloaded from the GitHub repository https://github.com/FedericaGalli17/Research-Assessment.

Overall, the database concerns 890 Italian Departments, 369 SDSs and 77965 products. Starting from this database, we selected those Departments that presented at least 50 products, ending up with 72756 products presented by 699 Departments. For each sector s of Department d, available information concerns the total number of products presented (P_{sd}), the percentage of products in each class of merit (ranging from A to F), the total score (v_{sd}), the mean score ($\bar{v}_{sd}$) and other quali-quantitative indicators used to assign the FFO performance-based share. The total score v_{sd} is calculated by multiplying the number of products linked to each category by their respective grades. Specifically, a grade of class A (excellent and extremely relevant) corresponds to a score of 1, class B (excellent) to 0.7, class C (standard) to 0.4, class D (sufficient) to 0.1, and classes E (poor relevance or not acceptable) and F (missing products) to a score of 0.

Average scores by SDS at the national level, $\bar{v}_s$, are reported in decreasing order in Fig. 1. Figure 2 reports the same average scores differentiating by Area.

Similar graphs are presented in Figure B1 and B2 of Appendix B for VQR 2015-2019 data, showing that the heterogeneity in the scores among the different sectors is preserved along the two research assessments. Although Fig. 1 highlights relevant differences in the outcome of the 2011-2014 research evaluation exercise among sectors, results are even more impressive across Areas: Area 10 is composed of several SDS with highly variable scores between and within sectors, Area 3 comprises few high-performing sectors, Area 8a received mean scores that are all below the national average level, in Areas 11b, 12 and 13 only one sector is exceeding the national average grade, while scores in Area 2, except for one sector, are highly polarized. Overall, as already noted by Baccini (2014a); De Nicolao (2014b); Baccini et al. (2020), discipinarly sectors evaluated through informed peer review (Areas from 10 to 14) are more likely to obtain low scores compared to sectors with high percentages of products assessed through bibliometric criteria (Areas from 1 to 9). This insight may depend both on the lower quality level of products belonging to these Areas as

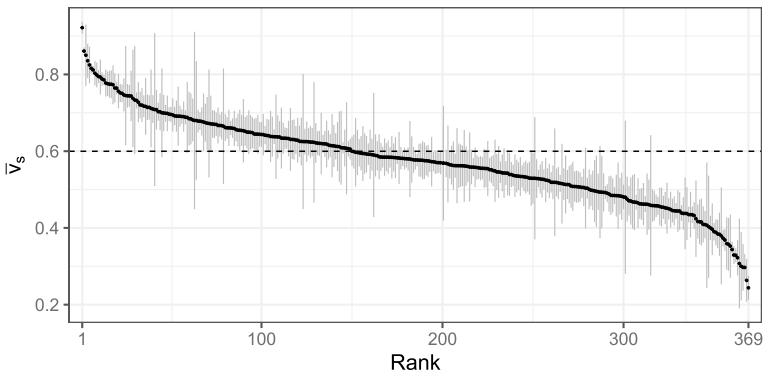

Fig. 1 SDSs average scores in decreasing order. Vertical segments (in grey) are obtained as $\bar{v}_s \pm 2\sigma_s$, where σ_s denotes the SDS-specific standard deviation. The dashed horizontal line corresponds to the global mean

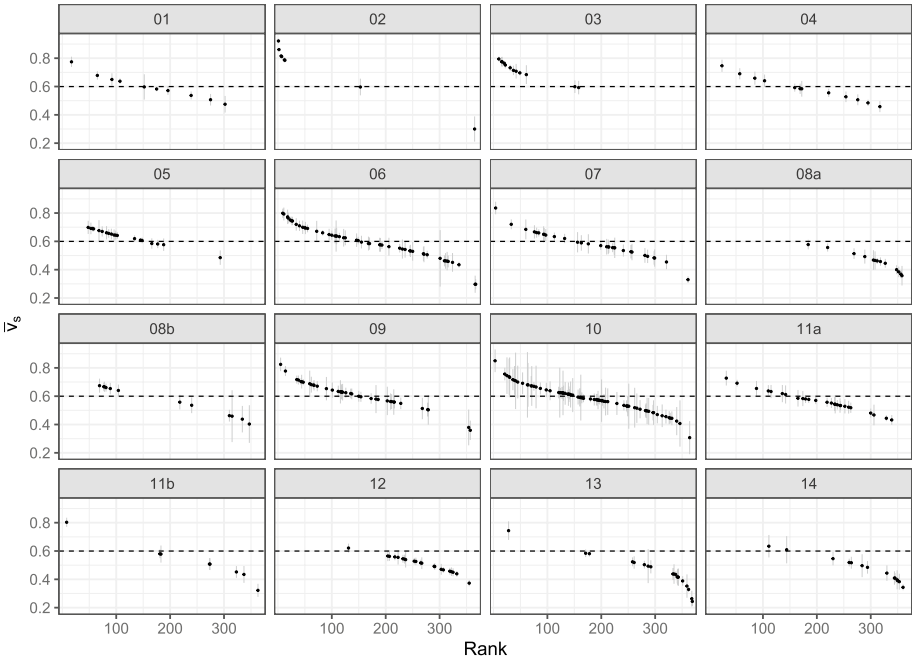

Fig. 2 SDSs average scores in decreasing order by Area

well as on the typology of the instruments used for the evaluation. Thus, relevant difficulties may arise in meaningfully interpreting the scores across Areas due to the impossibility of disentangling the two phenomena (Baccini & De Nicolao, 2016).

As further descriptive statistics, in Table 1 we provide the mean scores and the SDS-wise variance decomposition by Area. The overall mean score is 0.60, with a minimum value of 0.430 for Area 14 and a maximum of 0.818 for Area 2. The variance equals 0.122 with the highest variability in Area 13 (0.167) and the lowest in Area 12 (0.087). By decomposing the variance across Areas, we find that, overall, the scores are more variable within (86.9%) than across (13.1%) Areas. Thus, groups of experts that graded the products of a given Area assigned very heterogenous scores in particular in Areas 1, 5, 8b and 9, in which the within variability exceeds the 95% of the total variance. On the other hand, between variation exceeds 15% only in Areas 2, 3, 8a, 12, 13 and 14 as the scores are quite polarized among very high (Areas 2 and 3) or low values (Areas 8a, 12, 13, 14).

In sum, large differences in scores are detected both across and within Areas suggesting some difficulties in fairly assessing Departments' research quality due to the different possible reasons behind such heterogeneity (experts' evaluations, product selection, assessment procedure, etc...).

3 Indicators for Italian Deparments ranking

The methodological notes ANVUR (2017b) and ANVUR (2022b) start with the following statement on the logic underlying the ISPD indicator:

Table 1 SDS-wise Variance Decomposition by Area

Area	Average Score	Variance	Between Variance(%)	Within Variance(%)
01	0.60	0.138	2.7	97.3
02	0.82	0.121	42.0	58.0
03	0.77	0.115	18.2	81.8
04	0.56	0.127	5.7	94.3
05	0.64	0.120	2.5	97.5
06	0.60	0.140	9.1	90.9
07	0.57	0.124	9.2	90.8
08a	0.46	0.106	23.3	76.7
08b	0.62	0.126	4.5	95.5
09	0.64	0.125	4.8	95.2
10	0.57	0.091	5.5	94.5
11a	0.55	0.090	8.3	91.7
11b	0.54	0.145	13.6	86.4
12	0.49	0.088	16.3	83.7
13	0.46	0.167	18.6	81.4
14	0.43	0.120	29.2	70.8

"*The value of ISPD assigned to a specific department will be determined based on its positioning within the class of all departments with the same disciplinary composition (i.e., the same structure in terms of SSD of the members) that can be composed by permuting the staff present in those SSDs in Italian public institutions (Virtual Departments). ISPD will be obtained through the direct comparison among possible departments with the same disciplinary composition, and this comparison will be made in terms of the degree of success in the last VQR.*"

While the methodological note suggests that Virtual Departments (VDs) are derived by permuting the Scientific Departments' *affiliates*, in practical implementation, the permutation process is applied to the *scientific products* presented by these affiliates. Notably, the ISPD formulas rely on scores assigned to scientific products, and in all tables published in ANVUR reports, the focus is on products rather than affiliates. Despite this, we believe that establishing a departmental performance metric based on a comparison with Virtual Departments (VDs) is a highly effective approach for addressing the heterogeneity observed in Italian Departments with respect to SDS-wise composition.

Consider a Department d whose members are affiliated to a number S_d of SDSs and let $n_{sd},\ s = 1, \dots, S_d$, be the number of products presented by the affiliates of SDS s in Department d. Moreover, let $n_s,\ s = 1, \dots, S$, denote the number of products presented by affiliates of the SDS s at the national level. It turns out that the number of VDs associated with d is:

$$N_{\mathcal{V}_d} = \prod_{s=1}^{S_d} \binom{n_s}{n_{sd}} = \prod_{s=1}^{S_d} \frac{n_s!}{n_{sd}!(n_s - n_{sd})!},$$

where $\mathcal{V}_d$ denotes the set of all VDs associated to d, whose members will be denoted as $\mathcal{V}_d^{(k)},\ k = 1, \dots, N_{\mathcal{V}_d}$.

The comparison between a given Department and its associated VDs in terms of degree of success in the VQR requires to build a performance indicator obtained as a function of the scores assigned to each scientific product. Let v_{psd} be the score obtained by a scientific product p presented by a member of the Department d affiliated to the SDS s and collect such scores in the vector

$$\mathbf{v}_d = (v_{psd}), \;\; p = 1, \ldots, P_{sd}; \;\; s = 1, \ldots, S_d.$$

Let $\mathcal{I}(d)$ be a generic performance indicator for the Department d based on the VQR results, such that:

$$\mathcal{I}(d) = f(\mathbf{v}_d). \tag{1}$$

In principle, the procedure envisioned in ANVUR's methodological note requires computing the same indicator on the set $\mathcal{V}_d$:

$$\mathcal{I}\left(\mathcal{V}_d^{(k)}\right), \;\; k = 1, \ldots, N_{\mathcal{V}_d},$$

in order to obtain the standardized indicator:

$$\hat{\mathcal{I}}(d) = \frac{1}{N_{\mathcal{V}_d}} \sum_{k=1}^{N_{\mathcal{V}_d}} \mathbb{1}\left(\mathcal{I}(d) > \mathcal{I}\left(\mathcal{V}_d^{(k)}\right)\right); \qquad \hat{\mathcal{I}}(d) \in [0, 1] \tag{2}$$

where $\mathbb{1}$ denotes the indicator function. This measure corresponds to the probability that a random sample drawn from the national population of researchers in Italian institutions with the same SDS-wise composition of d shows a performance indicator $\mathcal{I}$ lower than that observed in Department d. Hence, the higher the value of such indicator, the better the performance of Department d. To rank Italian Departments, specifying the function $f(\cdot)$ in Equation (1), along with a method for computing the probability reported in Equation (2) is needed. Concerning the latter issue, since $N_{\mathcal{V}_d}$ is typically a huge number, it is unfeasible to compute the value of the indicator for all VDs, while exact calculation would involve the theory concerning the distribution functions of Multinomial random variables. However, such probability can be easily computed via Monte Carlo approximation or, when possible, resorting to the Gaussian approximation. Note that Monte Carlo approximations can be made arbitrarily accurate by increasing the number of Monte Carlo samples from the target random variable. Then, specifiying the function $f(\cdot)$ is by no means the most relevant issue to be tackled mainly because, as discussed in Sect. 2, the scores assigned to the scientific products in VQR can hardly be considered comparable between SDSs. The ISPD indicator proposed by ANVUR is only one of the possible solutions to the problem.

In what follows, after a description of the procedure followed to obtain the ISPD indicator, we study the effect of ANVUR's standardization of SDS scores and we address the statistical implications of such standardization on the Italian Departments ranking. Then, we propose some alternative indicators, in order to highlight important drawbacks of the procedure used in the last two selections of the Departments of Excellence: to this aim, we will also consider indicators based on different assumptions concerning SDS scores comparability.

3.1 ISPD

The construction of the ISPD index, as described in ANVUR methodological notes, starts with the computation of the standardized scores. The scores v_{psd} are standardized SDS-wise as:

$$\tilde{v}_{psd} = \frac{v_{psd} - \bar{v}_s}{\sigma_s},$$

where $\bar{v}_s$ and σ_s denote the mean and the standard deviation of the scores of the SDS s.

The Department-level indicator which corresponds to ANVUR's choice for Equation (1), corresponds to

$$\tilde{v}_d = f(\mathbf{v}_d) = \frac{1}{\sqrt{P_d}} \sum_{s=1}^{S_d} \sum_{p=1}^{P_{sd}} \frac{v_{psd} - \bar{v}_s}{\sigma_s} = \frac{1}{\sqrt{P_d}} \sum_{s=1}^{S_d} \sum_{p=1}^{P_{sd}} \tilde{v}_{psd}, \tag{3}$$

where P_d is the total number of products presented by Department d.

Concerning ANVUR's choice for Equation (2), the probability that a random sample drawn from the national population of researchers in Italian institutions with the same SDS-wise composition of d shows a performance indicator (3) lower than $\tilde{v}_d$ is computed by means of a Gaussian approximation, obtaining:

$$\text{ISPD}_d = 100 \times \Phi(\tilde{v}_d), \tag{4}$$

where $\Phi(\tilde{v}_d) = \Pr\left(\text{ISPD}\left(\mathcal{V}_d\right) < \tilde{v}_d\right)$ represents the cumulative distribution function of the standard Normal random variable, $\text{ISPD}\left(\mathcal{V}_d\right)$ is the ISPD in virtual Departments, and the factor 100 is used to scale the index range in the interval [0, 100].

3.2 Some remarks on the ISPD statistical properties

Due to its relevance in the funding of Italian Departments, the ISPD has received considerable attention in the debate concerning the quantitative assessment of the Italian Academic System: a prominent example of this debate can be found in De Nicolao (2014a), which is particularly interesting because it feats the leading researcher in the ISPD construction. In particular, the Gaussian approximation and scores' standardization (and subsequent aggregation) adopted by ANVUR raised several doubts. For instance, Bertoli-Barsotti (2017) first pointed out the main flows related to standardization while De Nicolao (2013) highlighted the difficulty for big Departments to achieve good ISPD values. In what follows, we start with some remarks on these topics, highlighting some drawbacks related to standardization. Nonetheless, in our opinion, the main flaw associated with ISPD is its tendency to polarization, which has important consequences on the ability of such indicator to effectively discriminate between outstanding Departments as opposed to those just above the average: eventually, it will be shown that this has relevant implications on funding allocation.

3.2.1 Gaussian approximation

The adequacy of the Gaussian approximation adopted by ANVUR in Equation (4) can be checked following a Monte Carlo approach: the computation of the sought probability requires a B-dimensional sample of VDs, with B sufficiently large in order to guarantee accuracy. For each sampled VD, the standardized score $\tilde{v}\left(\mathcal{V}_d^{(b)}\right)$, $b = 1, \ldots, B$, is computed and the Monte Carlo computation of ISPD corresponds to

$$\text{ISPD}_d^{\text{MC}} = \frac{100}{B} \sum_{b=1}^{B} \mathbb{1}\left(\tilde{v}_d > \tilde{v}\left(\mathcal{V}_d^{(b)}\right)\right) \tag{5}$$

Such an approximation is justified by the Central Limit Theorem (CLT): since $\tilde{v}\left(\mathcal{V}_d^{(b)}\right)$ consists in a sum of many standardized variables, CLT ensures that its distribution can be approximated by a Gaussian distribution, provided that the sample size (P_d) is sufficiently large. In its more classic formulation, CLT is stated under the assumption of independence; yet, extensions of the CLT ensure its validity, at a lower convergence rate, even if the independence assumption is relaxed. As a matter of fact, since by law 240/2010 each Department must be composed of no less than 40 members and in VQR 2011-2014 each member must present 2 products (in VQR 2015-2019 the number of products presented per person raises to 3), a sufficiently high sample size is guaranteed.

In fact, Fig. 3 displays the accuracy of the Gaussian approximation: results are based on $B = 500,000$ Monte Carlo samples, using VQR 2011-2014 data analyzed in this paper. The left panel shows that the Standard Normal density accurately fits the distribution of standardized scores in VDs associated with a given Department. The middle and right panels of Fig. 3 show the agreement between ISPD_d and $\text{ISPD}_d^{\text{MC}}$ for all the considered Departments: the discrepancy is in general lower than 0.5, highlighting that the approximation is accurate up to the third digit.

3.2.2 Standardization

The primary drawback of standardization lies in its inherent nature: standardized scores eliminate the between-SDS component of the score variance. In a comparative evaluation

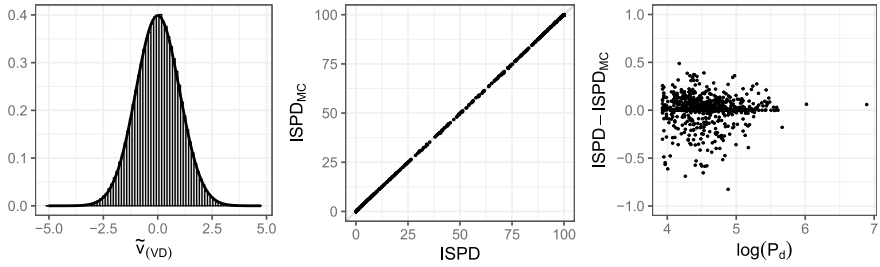

Fig. 3 Left panel: distribution of standardized scores in VDs and Gaussian approximation for an example Department. Middle panel: comparison between ISPD based on Gaussian approximation and ISPD obtained via Monte Carlo sampling (ISPD_{MC}). Right panel: differences ISPD − ISPD_{MC} as a function of the log number of products presented by Departments

geared towards recognizing excellence, accepting this outcome is only reasonable if the assumption that there are no disparities in the research quality among SDSs holds. This assumption further presupposes that all observed deviations from average scores stem from divergent attitudes of evaluation committees and various unique features inherent to each SDS, such as evaluation methods and typical publication channels. For insights into these distinctive characteristics, refer to ANVUR's response to the queries raised by the Italian National University Council regarding the ISPD indicator (ANVUR, 2017a). In a (more realistic) scenario where some SDSs exhibit research of high international relevance while other SDSs demonstrate lower quality, the employment of standardized scores would overlook the distinctions in quality between SDSs. Moreover, this approach could be detrimental, as it might penalize high-quality sectors in comparison to their low-quality counterparts.

However, deliberating on the comparability of scores between SDSs and the disparities in research quality among SDSs, which would necessitate comparisons between the scientific output of Italian and international researchers, falls beyond the scope of this paper. Our focus here is solely to provide some insights into the coherence of the entire VQR process concerning construction of statistical indicators adopted for funding allocation. It is essential to highlight that the indicators utilized for the allocation of the FFO performance-based share operate under the assumption that SDS scores are comparable within Scientific Areas but not between Scientific Areas, as elaborated further in Sect. 3.3.1. This stands in contrast to the global incomparability assumed in constructing the ISPD.

From a practical standpoint, standardization introduces a situation where a scientific product belonging to an SDS with a higher average score will yield lower standardized scores than products with an equivalent score but presented by affiliates of an SDS showing a lower average score. To illustrate this, refer to Table 2, where it is shown that two sectors with a mean score of 1 within a Department (all products graded as A) achieve different mean standardized scores ($\tilde{v}_{sd}$) owing to the diverse average scores of their associated national-level SDSs. Specifically, SDS-1 outperforms SDS-2 since $\overline{v}_{\text{SDS-1}} < \overline{v}_{\text{SDS-2}}$. Additionally, while SDS-3 obtains a negative mean standardized score (−0.18), SDS-4 is graded positively (0.18), despite both having the same mean score of 0.5, due to $\overline{v}_{\text{SDS-3}} > \overline{v}_{\text{SDS-4}}$. The last five columns of Table 2 show standardized scores referred to categories A-E. It can be noticed that, because of the differences in $\overline{v}_s$, products classified as A in SDS-4 (the worst performing SDS included in the Table) receive the highest score, while products classified as E in SDS-2 (the best performing SDS included in the Table) receive the lowest score.

In essence, while standardization aims to eradicate potential discrepancies in evaluations resulting from diverse approaches employed by the experts in various fields, it

Table 2 Practical implications of standardization

	Department d		All Italian institutions						
	$\overline{v}_{sd}$	$\tilde{v}_{sd}$	$\overline{v}_s$	σ_s	$\tilde{v}_s^A$	$\tilde{v}_s^B$	$\tilde{v}_s^C$	$\tilde{v}_s^D$	$\tilde{v}_s^E$
SDS-1	1	1.06	0.66	0.32	1.06	0.12	− 0.81	− 1.75	− 2.06
SDS-2	1	0.88	0.72	0.32	0.88	− 0.06	− 1.01	− 1.96	− 2.27
SDS-3	0.5	− 0.18	0.56	0.34	1.28	0.40	− 0.47	− 1.35	− 1.64
SDS-4	0.5	0.18	0.44	0.34	1.65	0.77	− 0.11	− 0.99	− 1.29

inadvertently skews product scores of different SDSs, as evidenced in Table 2. Simultaneously, it hinders the assessment and comparison of research quality between SDSs.

3.2.3 Dependence on the sample size and tendency to polarization

Note that the standardized score $\tilde{v}_d$ defined in Equation (3) can be expressed as the product $\bar{x}_d\sqrt{P_d}$, where

$$\bar{x}_d = \frac{1}{P_d}\sum_{s=1}^{S_d}\sum_{p=1}^{P_{sd}}\tilde{v}_{psd}, \tag{6}$$

i.e., as the product of an average measure of the quality of the Department scientific production as measured by the average standardized scores ($\bar{x}_d$), times a measure of the Department size ($\sqrt{P_d}$).

By multiplying $\bar{x}_d$ for the positive constant $\sqrt{P_d}$, the values of the standardized score $\tilde{v}_d$, and consequently of the ISPD, become more extreme the greater the number of products presented by the Department d.

This is shown in Fig. 4 where small Departments ($P_d < 80$) are compared with larger Departments ($P_d > 150$). As expected from standard statistical theory, larger Departments show less variability than small Departments with respect to the distribution of $\bar{x}_d$, as variability of the sample mean is inversely proportional to the sample size. On the contrary, passing from $\bar{x}_d$ to $\tilde{v}_d$ leads to a relevant stretch in the empirical cumulative distribution function for larger Departments, that shows more extreme results. In sum, if $\bar{x}_d > 0$, as P_d increases, the values of the ISPD will be closer and closer to 100, while if $\bar{x}_d < 0$ the ISPD will approach 0 for increasing values of P_d. As an example, consider two Departments with the same average score $\bar{x}_d = 0.1$, but with different sizes, say $P_1 = 80$ and $P_2 = 240$. It turns out that $\tilde{v}_1 = 0.89$ and $\tilde{v}_2 = 1.55$, delivering $\text{ISPD}_1 = 81.5$ and $\text{ISPD}_2 = 94$. The difference between the ISPD scores is totally ascribable to the Department sizes, as the average score that, following the phrasing of ANVUR's methodological note should measure the degree of success in the VQR exercise, is equal. Note that the first Department needs to obtain an average score $\bar{x}_1 = 0.173$ in order to get the same ISPD as the larger Department, i.e. it needs to almost double its performances in terms of research quality.

Dependency on the size of the Departments leads to the most relevant flaw of the ISPD indicator: its tendency to polarization. To see this, it is useful to interpret the ISPD as a p-value arising from a statistical test, recalling that a p-value can be interpreted as the probability to observe a more extreme result than the observed one, if the null hypothesis is true. As noticed at the end of Sect. 3.1, the random variable describing standardized scores within VDs associated with Department d, is well approximated by a standard Normal distribution. Let $\overline{X}_{\mathcal{V}_d}$ be the random variable describing the average score in VDs. One gets:

$$\overline{X}_{\mathcal{V}_d} \sim N\left(0, \frac{1}{P_d}\right), \tag{7}$$

so that the random variable referred to standardized scores in VDs

$$\tilde{V}_{\mathcal{V}_d} = \sqrt{P_d}\,\overline{X}_{\mathcal{V}_d} = \frac{\overline{X}_{\mathcal{V}_d} - 0}{1/\sqrt{P_d}} \sim N(0,1) \tag{8}$$

corresponds to the usual test statistic employed to test the hypothesis system where the null hypothesis states that the score of Department d is equal to that of VDs, which is zero by construction, while the alternative hypothesis states that the score of Department d is greater than zero. In this spirit, the Department standardized score $\tilde{v}_d$ corresponds to the observed value of the test statistic (8) and the ISPD can be obtained as:

$$\text{ISPD}_d = 100 \times \left[1 - \Pr\left(\tilde{V}_{\mathcal{V}_d} > \tilde{v}_d | \text{H}_0\right)\right], \tag{9}$$

i.e. 1 minus the p-value associated to the aforementioned test. The polarization issue characterizing ISPD arises as a well-known feature of the frequentist tests: for large sample sizes, the p-value shows extreme values also when differences between the sample mean and the value specified under the null hypothesis are very small and possibly irrelevant from a practical point of view. Therefore, ISPD values can become extremely high (or low) in case of small positive (or negative) deviations from the mean. This is self-evident if one looks at the ISPD values published by ANVUR for the Departments of Excellence selection in both VQR exercises, as will be thoroughly discussed in Sect. 4.

3.3 Some alternative indicators

In this section, we explore alternative approaches to the ISPD indicator, emphasizing that these proposals serve as analytical tools for comprehending the principal drawbacks associated with the ISPD through comparisons. They are not intended as actual suggestions for new indicators to be adopted in the selection of Departments of Excellence. Specifically, in Sect. 3.3.1, we delve into indicators that align with the rationale guiding the FFO performance-based share allocation, adapting them to a departmental context. Implementing such indicators for ranking Italian Departments would enhance the coherence between the criteria used in the selection of Departments of Excellence and the distribution of the FFO performance-based share. Currently, a notable inconsistency lies in the assumptions regarding the comparability of SDS-specific scores. In the selection of Departments of Excellence, scores are deemed globally incomparable and standardized within each SDS. In contrast, FFO performance-based share allocation relies on indicators constructed under the assumption that SDS scores are comparable within Scientific Areas. For a detailed description of these indicators, refer to, for instance, ANVUR (2022a). We advocate for the desirability of such coherence, given that both tasks

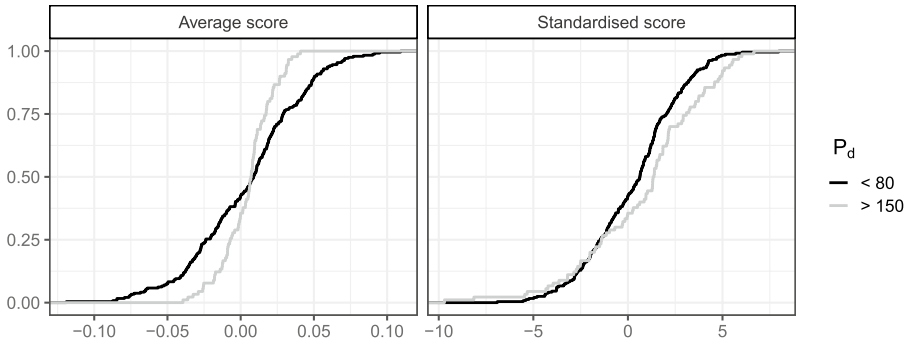

Fig. 4 Empirical cumulative distribution function of average scores $\bar{x}_d$ (left panel) and standardized scores $\tilde{v}_d$ (right panel)

utilize the same VQR results. Simply put, what is deemed comparable or incomparable for one task should be approached in the same manner for the other task.

Finally, in Sect. 3.3.2 we propose an indicator that solves some of the critical aspects of standardization adopted by the ISPD, but that works fully in the spirit of ISPD when the idea of comparison with virtual Departments is considered.

3.3.1 Indicators grounded on the FFO performance-based share allocation

FFO performance-based share allocation is carried out by means of quali-quantitative indicators taking into account both a measure of quality and a measure of the size of the institution under evaluation. In this Section we focus on the indicator measuring institutions' quality, known as the R index on which the quali-quantitative indicators used to allocated the FFO performance-based share are built.

The baseline R indicator used by ANVUR to allocate the FFO performance-based share, as established by Bando VQR 2011-2014 (see ANVUR (2022a)), is based on the comparison between the SDS average score and the average score of the Scientific Area to which the SDS belongs, that we denote as $\overline{v}_s^a$, so that the Department level indicator can be expressed as

$$R_d = \sum_{s=1}^{S_d} \frac{P_{sd}}{P_d} \frac{\overline{v}_{sd}}{\overline{v}_s^a}. \tag{10}$$

Here, we also adapt the R index at the Departmental level with the important modification that the index reported in Equation (11) assumes that SDS scores are not comparable between SDSs, as hypothesized in the construction of the ISPD indicator. Hence, the index can be computed by dividing the mean score of each SDS s in the Department d ($\overline{v}_{sd}$) by the overall mean score of the sector s ($\overline{v}_s$) as

$$R^*_{sd} = \frac{\overline{v}_{sd}}{\overline{v}_s},$$

so that, if the Department d in the sector s shows a higher score than the average score of the sector s, $R^*_{sd} > 1$. Finally, the Department level indicator can be computed by weighted average as

$$R^*_d = \sum_{s=1}^{S_d} \frac{P_{sd}}{P_d} R^*_{sd}. \tag{11}$$

Indicators R and R^* will be compared in Sect. 4 to shed some light on the drawbacks of the ISPD. Moreover, we will consider the mean score by sector and Department, commonly indicated as "I" in the VQR reports, that is obtained as

$$\overline{v}_d = \frac{1}{P_d} \sum_{s=1}^{S_d} \sum_{p=1}^{P_{sd}} v_{psd}. \tag{12}$$

Note that this indicator assumes full comparability among SDS scores and it is not used by ANVUR for any allocation task. Yet, we find it interesting to compare the distribution of this index with other indicators considered in this paper.

3.3.2 Standardization-free ISPD

In this Section, we propose an alternative indicator of departmental performance that overcomes the criticisms due to the sum of standardized scores characterizing ANVUR's ISPD. The standardization-free indicator, denoted as ISPD^*_d, starts by computing SDS-specific probabilities

$$\Pr\left(\tilde{v}_{sd} > \tilde{v}\left(\mathcal{V}_{sd}\right)\right), \qquad s = 1, \ldots, S_d$$

i.e. the probability that the score of the affiliates of SDS s in Department d, $\tilde{v}_{sd}$, is higher than the score, $\tilde{v}\left(\mathcal{V}_{sd}\right)$, of a P_{sd}-dimensional random sample from the whole population of the affiliates of s in all the Italian institutions, where P_{sd} is the number of products of the SDS s of Department d. In this spirit, $\mathcal{V}_{sd}$ denotes the set of Virtual Departments associated to Department d for SDS s.

Note that $\Pr\left(\tilde{v}_{sd} > \tilde{v}\left(\mathcal{V}_{sd}\right)\right) = \Pr\left(v_{sd} > v\left(\mathcal{V}_{sd}\right)\right)$, since

$$\tilde{v}_{sd} = \sum_{p=1}^{P_{sd}} \frac{v_{psd} - \overline{v}_s}{\sigma_s} > \tilde{v}\left(\mathcal{V}_{sd}\right) = \sum_{p=1}^{P_{sd}} \frac{v_p\left(\mathcal{V}_{sd}\right) - \overline{v}_s}{\sigma_s} \implies \sum_{p=1}^{P_{sd}} v_{psd} > \sum_{p=1}^{P_{sd}} v_p\left(\mathcal{V}_{sd}\right) \implies v_{sd} > v\left(\mathcal{V}_{sd}\right)$$

hence the same probability is obtained by using standardized or non-standardized scores.

The computation of the sought probabilities is performed with the same logic underlying the computation of ISPD^{MC} in Equation (5), delivering the SDS-specific indicator:

$$\text{ISPD}^*_{sd} = \frac{100}{B} \sum_{b=1}^{B} \mathbb{1}\left(\tilde{v}_{sd} > \tilde{v}\left(\mathcal{V}^{(b)}_{sd}\right)\right); \qquad s = 1, \ldots, S_d.$$

Finally, the standardization-free ISPD for Department d can be computed as a weighted average of SDS-specific indicators:

$$\text{ISPD}^*_d = \frac{1}{P_d} \sum_{s=1}^{S_d} P_{sd} \times \text{ISPD}^*_{sd}, \tag{13}$$

where P_d is the number of products presented by the affiliates of Department d. Note that this indicator consists in a weighted average of within-SDS rankings, so that the contribution of each SDS in a Department arises from a fair comparison with researchers in the same SDS: we believe that this improves the pitfalls of ISPD related to the summation of standardised scores. Because of the small number of products by sector and Department, the Gaussian approximation adopted for computing the ISPD cannot be employed, hence, the ISPD^* computation requires Monte Carlo sampling. While this could be time demanding, we believe that the computational burden is perfectly bearable when dealing with indicators that will be used to allocate huge amounts of funding. A non-optimized R code took 6.5 h to obtain ISPD^* for the 699 Departments considered in our application on a laptop equipped with 32Gb RAM and 12 core CPU.

The most important difference between ISPD and ISPD^* is constituted by the methodology adopted to aggregate SDS-specific performances: while ISPD aggregates standardized scores in order to obtain $\tilde{v}_d$ that are compared with VDs afterwards, ISPD^* aggregates SDS-specific probabilities that can be interpreted as the rank of affiliates of

SDS s of Department d among all possible groups of the same size that can be formed from the whole population of SDS s affiliates in Italy. This overcomes the criticisms due to standardization discussed in Sect. 3.2. In our opinion, from a statistical point of view, a weighted sum of probabilities is way less controversial than the sum of standardized scores. Moreover, because of its structure, $ISPD_d^*$ will show less sensitivity to the Department size as measured by P_d, solving the aforementioned polarization issue, as shown in Sect. 4.

To sum up, Table 3 reports the assumptions on SDS score comparability characterizing the considered indices. Despite the frequent invocation of the principle that indicators should be constructed for a specific purpose and only used for that purpose, we believe that assumptions regarding the comparability of scores from different SDS should be taken into account at the outset of indicator construction and should be consistent across all exercises aimed at allocating funds.

4 Empirical application

In this Section, we illustrate the ISPD issues discussed in Sect. 3 through empirical evidence, utilizing data from the Italian research assessment VQR 2011-2014. In particular, we focus on the 2011-2014 data because, although more recent VQR results are available, the 2015-2019 reports lack disaggregated data at the SDS and Department levels, preventing us from calculating the aforementioned indicators.

Figure 5 shows the distribution of the scores obtained by the 699 Italian Departments considered in the analysis differentiating among ISPD, ISPD , R, R and I. While the distribution appears to be bell-shaped according to the indicators used to distribute the FFO performance-based share and ISPD , the ISPD values, as expected, are highly polarized. Most of the Departments are graded as "extremely bad" or "extremely good" by the ISPD: 97 and 150 Departments of out 699 reach a score equal to 0 and 100, respectively. The majority of Departments reach extreme ISPD values (167 between 0 and 5, and 215 between 95 and 100) with only 317 ranging in the middle. On the other hand, much more variability is given by the ISPD , R, R and I classification.

These insights are corroborated by the data presented in Fig. 6. In both panels, the left y-axis illustrates Departments with ISPD scores below 5 and above 95. The scores corresponding to these Departments in terms of ISPD (depicted in the left panel) and R index (depicted in the right panel), are indicated on the right y-axis. It is evident that both ISPD and R indices lead to a notably higher degree of disparity in the Departments scores. Therefore, while it is difficult to draw a ranking of the best Italian Departments based on ISPD since most of them reach similar results, indicators such as the ISPD allow obtaining wider differences in the scores across Departments that can be useful in establishing a ranking.

In summary, we believe that the ISPD indicator is not suitable for building a ranking able to identify excellent Departments. In reality, ISPD merely distinguishes between Departments above and below the average. Consequently, Departments demonstrating outstanding scientific production receive comparable scores to those slightly above the average, undermining the notion of funding based on excellence, as illustrated in the subsequent discussion.

Table 3 Comparability between SDS scores assumed by each indicator considered in Sect. 4

Assumption	ISPD_d	ISPD_d^*	R_d^*	R_d	$\overline{v}_d$
Not comparable overall	+	+	+	−	−
Comparable within Areas	−	−	−	+	−
Comparable overall	−	−	−	−	+

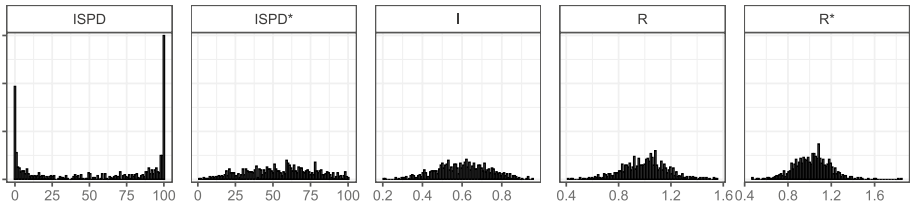

Fig. 5 Distribution of the scores based on the different indicators

4.1 Implications on funding

In this section, we discuss how the use of the ISPD indicator to rank Departments may affect the Departments of Excellence funding allocation.

The procedure followed to select the 180 Departments of Excellence can be outlined as follows. First, the best 350 Departments in terms of ISPD are selected based on the VQR results. Each University in the top 350 is then asked to present, for a maximum of 15 Departments, a five-year Departmental development project, accompanied by a financial program. In the first phase, for each University in the ranking, a commission of seven members named by ministerial decree evaluates the application presented by the Department that obtained the best placement in the top 350 and in case of favourable evaluation, this Department is selected for funding. For the remaining available positions, the selection is based on the final total score (S_{tot}) which is obtained as the sum of the score attributed to the project (S_{proj}, maximum 30 points) and the ISPD multiplied times 0.7 (S_{ISPD}), i.e.:

$$S_{\text{tot}} = 0.7 \times \text{ISPD} + S_{\text{proj}} = S_{\text{ISPD}} + S_{\text{proj}}$$

Table 4 shows some statistics on the ISPD and project score based on the available data on the 180 Departments of Excellence selected in the years 2018-2022 and 2023-2027. Note that, while the 2018-2022 Departments of Excellence selection refers to the VQR 2011-2014 results, the 2023-2027 selection is based on VQR 2015-2019.

The most remarkable difference between the 2018-2022 and the 2023-2027 selection concerns the number of products evaluated: while two products per person were evaluated in VQR 2011-2014 used to select the 2018-2022 Departments of Excellence, the number of products raised to three in the following selection (VQR 2015-2019). As a consequence, the tendency of the ISPD to reach extremely high values for well performing Departments is more evident in the period 2023-2027. Indeed, in 2023-2027 the mean ISPD score is even more close to 70 compared to the previous selection and the variance noticeably decreases with respect to the 2018-2022 period passing from 9.43 to 3.98. Due to the ISPD polarization tendency, as shown by the coefficient of variation (CV), most of the variability in the ranking of the 180 Departments of Excellence is explained by the project score rather than by the ISPD. Note that the correlation between S_{ISPD} and S_{proj} is equal to 0.26

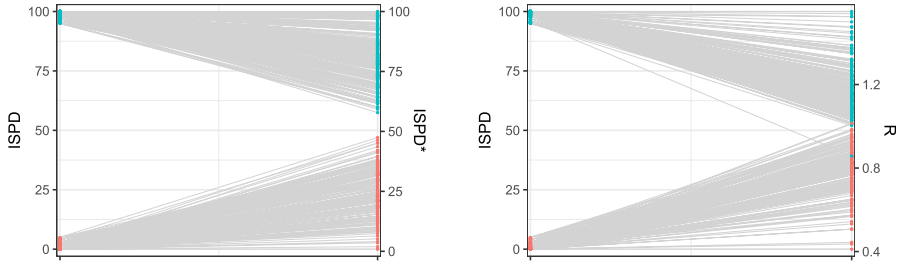

Fig. 6 ISPD polarization tendency. ISPD vs ISPD* (left panel) and ISPD vs R (right panel)

and 0.12 in the two periods. Therefore, given the low variability of the ISPD, the project presented by the 350 eligible Departments results to be the most important factor determining the final funds allocation.

These insights are outlined also in Fig. 7 which shows that while the ISPD values of the 180 Departments of Excellence are very concentrated around 100, most of the variability can be attributed to the project score. Note that a more insightful analysis of the role of the ISPD in determining the final ranking of the Departments of Excellence should have involved data on all the 350 eligible Departments which are actually not available. The only information at our disposal on all the top 350 Departments concerns the ISPD, while S_{proj} is officially published only for the 180 Departments of Excellence. Therefore, in the next paragraph, we concentrate on it.

Table 5 confirms previous findings on the extreme mean values reached by the ISPD among the top 350 Departments as well as on the greater tendency to polarization for increasing sample sizes. Indeed, the variance halves passing from two to three products presented per person and the coefficient of variation in 2023-2027 is about two-thirds of the CV in 2018-2022. Comparing these results to data on VQR 2011-2014 on the 350 best performing Departments based on the ISPD, we confirm previous insights on the high mean level and low variability of the ISPD.[4] Moreover, statistics on the alternative indicators show that much more variability is granted by ISPD , R, R and I, reaching CV values almost four or fives times bigger than the one of the ISPD. Therefore, as noted in the previous sections, these alternative indicators provide more suitable results to establish a ranking of the Italian Departments since their distribution is more variable across units and allow to discriminate between truly excellent Departments with respect to those slightly above the average.

5 Conclusion

Given the relevance of performance-based funding for Italian Departments and the heated debate in the Italian Academic System on research evaluation adoption and methods, in this paper, we analyse the statistical properties of the indicator used to establish a ranking

[4] As explained before, the VQR 2011-2014 data used in this paper do not allow to exactly reproduce ISPD values officially published when selecting the Departments of Excellence because of shifts in the composition of the Departments in time with respect to tables used in this work and of the non-exhaustiveness of the data provided by the tables available online due to confidentiality restrictions.

Table 4 ISPD and Project score of the 180 Departments of Excellence: summary statistics

	Mean		Var		CV	
	2018–2022	2023–2027	2018–2022	2023–2027	2018–2022	2023–2027
S_{ISPD}	68.4	69.2	9.43	3.98	0.045	0.029
S_{proj}	25.4	25.7	10.28	9.78	0.126	0.122
Covariance	–	–	2.61	0.75	–	–
S_{tot}	93.8	94.9	24.93	15.26	0.053	0.041

of the 350 best Italian Departments, i.e. the ISPD. Our analysis shows that the two main weaknesses of the ISPD concern its dependence on the standardization procedure and its tendency to polarization.

First, ISPD does not take SDS' research quality level into account because standardized scores have zero means and the between component of the variance is null for construction. Since the ISPD is based on standardized scores, the final ranking results can be considered as reliable only under the assumption that there are no differences in the research quality of each SDS. This appears as a questionable assumption that could be checked via comparison with researchers from non-Italian institutions.

The second, and likely the most relevant issue of the ISPD concerns its dependence on the positive constant $\sqrt{P_d}$ which generates more polarized scores the bigger the sample size is. This feature can be explained by considering the ISPD as a p-value in frequentist tests. Therefore, for large sample sizes, the ISPD values can become extremely high or low in case of small (positive or negative) deviations from the mean.

Finally, being ISPD an indicator at the Department level which cannot be decomposed at SDS level, it is not helpful in designing within Department policies. However, it would be extremely useful for Departments to implement ad hoc policies and plans targeting specific groups of researchers in order to further exploit existing opportunities and reduce possible weaknesses.

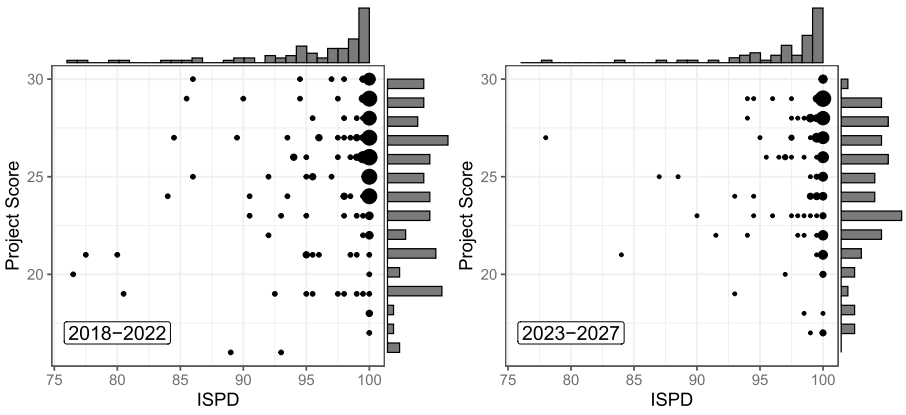

Fig. 7 ISPD vs Project score: 180 Departments of Excellence

Table 5 Indicators' summary statistics for the best 350 Departments

	VQR 2011-2014 data			Ranking best 350 Departments data					
				2018–2022			2023–2027		
	Mean	Var	CV	Mean	Var	CV	Mean	Var	CV
ISPD	97.49	11.86	0.035	93.28	81.49	0.097	96.02	40.60	0.066
ISPD*	78.15	141.70	0.152	–	–	–	–	–	–
R	1.18	0.023	0.128	–	–	–	–	–	–
R*	1.21	0.026	0.132	–	–	–	–	–	–
I	0.70	0.012	0.154	–	–	–	–	–	–

In order to overcome the issues of the ISPD, possible alternative indicators have been proposed such as a standardization-free ISPD computed based on SDS-specific probabilities and indices relying on the same logic as indicators used to allocate the FFO performance-based funds. Indeed, we believe that coherence between FFO performance-based funding and the Departments of Excellence selection is desirable as both tasks are achieved by using the same VQR results. As shown in the empirical section using the 2011-2014 VQR data, these alternative indicators allow to obtain more variable scores which is a desirable property for the purpose of ranking Departments. Indeed, the top 350 Italian Departments tend to have very high and similar ISPD values and thus, the final ranking mostly depends on the project score.

Since the specific aim of this paper is to solely discuss the statistical properties of the ISPD indicator, in the near future, more research should be devoted to analysing other critical aspects of the Italian research assessment exercise. Some relevant points that would need further investigation include (i) possible distortions arising from considering a fixed number of products per researcher both in sectors with high publishing rates as well as in less productive sectors; (ii) the role of some relevant confounding factors such as the North–South divide and the uneven infrastructure and transport network connecting the different Italian Universities in shaping departmental performance.

Given the great socio-economic impact of research evaluation exercises (for an examination of media coverage and its impact on Universities' visibility and reputation see Blasi et al. (2016, 2017)), we advocate for the use of more statistically robust indicators. Moreover, although research is one of the primary drivers of economic development, in Italy, non-competitive funding for research remains insufficient (Zanassi, 2015), placing excessive emphasis on evaluation exercises for funding allocation. Finally, in view of the great difficulties encountered in analysing official data due to limited data availability, in the future, we wish for more transparency of the results published by ANVUR for the sake of reproducibility.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-025-05240-2.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors have no Conflict of interest to declare.

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.

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*, 2053–2065.

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2014). Inefficiency in selecting products for submission to national research assessment exercises. *Scientometrics, 98*(3), 2069–2086.

Akbaritabar, A., Bravo, G., & Squazzoni, F. (2021). The impact of a national research assessment on the publications of sociologists in Italy. *Science and Public Policy, 48*(5), 662–678.

ANVUR (2017a). "L'indicatore standardizzato di performance dipartimentale (ISPD). Risposta al Comunicato CUN del 18 luglio 2017". In: pp. 1–3. https://www.anvur.it/wp-content/uploads/2017/10/RispostaANVURaCUNsuusiISP~.pdf.

ANVUR (2017b). "Nota metodologica sul calcolo dell'indicatore ISPD". In: pp. 1–3. https://www.anvur.it/wp-content/uploads/2018/04/Nota_metodologica_ISPD_AN_.pdf.

ANVUR (2022a). "Calcolo dei profili di qualità e degli indicatori per la VQR 2015-2019". In: https://www.anvur.it/wp-content/uploads/2022/04/Calcolo-dei-profili-di-qualita%CC%80-e-degli-indicatori-per-la-VQR-2015_2019.pdf.

ANVUR (2022b). "Nota metodologica approvata dal Consiglio Direttivo con delibera n. 93 del 27 aprile 2022". In: pp. 1–3. www.mur.gov.it/sites/default/files/2022-07/Nota%20metodologica%20ISPD%202022.pdf.

Baccini, A. (2014a). "La VQR di Area 13: una riflessione di sintesi". In: *Statistica & Societa* 3.3, pp. 32–37.

Baccini, A., Barabesi, L., & De Nicolao, G. (2020). On the agreement between bibliometrics and peer review: Evidence from the Italian research assessment exercises. *PLoS ONE, 15*(11), e0242520.

Baccini, A., & De Nicolao, G. (2016). Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. *Scientometrics, 108*, 1651–1671.

Baccini, A., & De Nicolao, G. (2022). Just an artifact? The concordance between peer review and bibliometrics in economics and statistics in the Italian research assessment exercise. *Quantitative Science Studies, 3*(1), 194–207.

Bertoli-Barsotti, L. (2017). *Le incongruenze dell'ISPD e i dipartimenti di eccellenza*. Ed. by ROARS. https://www.roars.it/le-incongruenze-dellispd-e-i-dipartimenti-di-eccellenza/.

Blasi, B., Romagnosi, S., & Bonaccorsi, A. (2016). Playing the ranking game: Media coverage of the evaluation of the quality of research in Italy. *Higher Education, 73*, 741–757.

Blasi, B., Romagnosi, S., & Bonaccorsi, A. (2017). Universities as celebrities? How the media select information from a large research assessment exercise. *Science and Public Policy, 45*(4), 503–514.

Bonaccorsi, A. (2020). Two decades of experience in research assessment in Italy. *Scholarly Assessment Reports, 2*(1), 1–19.

Checchi, D., Mazzotta, I., Momigliano, S., & Olivanti, F. (2020). Convergence or polarisation? The impact of research assessment exercises in the Italian case. *Scientometrics, 124*, 1439–1455.

De Nicolao, G. (2013). *L'ANVUR, la classifica degli atenei della VQR e la legge dell'imbuto*. Ed. by ROARS. https://www.roars.it/lanvur-la-classifica-degli-atenei-della-vqr-e-la-legge-dellimbuto/.

De Nicolao, G. (2014). *Voti VQR ai dipartimenti: le normalizzazioni fai-da-te di CRUI e ANVUR*. Ed. by ROARS. http://www.https://www.roars.it/voti-vqr-ai-dipartimenti-le-normalizzazioni-fai-da-te-di-crui-e-anvur/.

De Nicolao, G. (2014). *VQR da buttare? Persino ANVUR cestina i voti usati per l'assegnazione FFO 2013*. Ed. by ROARS. http://www.roars.it/online/vqr-da-buttare-persino-anvur-cestina-i-voti-usati-per-lassegnazione-ffo-2013/.

Demetrescu, C., Ribichini, A., & Schaerf, M. (2020). Are Italian research assessment exercises size-biased? *Scientometrics, 125*, 533–549.

Marzolla, M. (2015). Quantitative analysis of the Italian National Scientific Qualification. *Journal of Informetrics, 9*(2), 285–316.

Poggi, G. (2015). L'esercizio VQR e la valutazione dei dipartimenti. In G. Conte (Ed.), *Evoluzione e valutazione della ricerca giuridica.* ESI.

Poggi, G., & Nappi, C. A. (2014). Il Voto standardizzato per l'esercizio VQR 2004–2010. *RIV Rassegna Italiana di Valutazione, 59*, 34–58.

Spanò, R., Bracci, E., Manes-Rossi, F., Sforza, V. (2024). "A fatally efficient machine. Insights into the 'banality' of the research evaluation exercise in Italy". In: *Critical Perspectives on Accounting* 100, p. 102742.

Zanassi, S. (2015). Evaluating and financing research: A comparison among universities in Italy, France, Germany and Spain. *Italian Journal of Educational Research, 15*, 1–16.