The effect of research evaluation exercises on research output: Fifteen years of evidence from Italy*

Massimiliano Bratti[a], Tindaro Cicero[b], Enrico Lippo[a], Carmela Anna Nappi[b], Matteo Turri[c]

This is the version firstly submitted to the journal *Politica Economica - Journal of Economic Policy* and different from the final published version.

Abstract

The purpose of this paper is to investigate the potential benefits from implementing Performance-based University Research Funding Systems (PRFSs) in terms of both quantity and quality of research output measured in the Web of Science database. The Italian experience of research evaluation, in particular two PRFSs carried out in the 2000-2015 period, is examined. We use a *Difference-in-Differences in Reverse* strategy, in which a country --- the UK --- where PRFSs were always in place is compared with a country that switched from not having to implementing PRFSs --- Italy. Our analysis confirms that PRFSs are associated with an improvement in the "average quality" of research output (% of documents cited), although they do not increase outputquantity (number of articles) or "excellence" (measured as the percentage of top cited articles). These effects are not generalized though, and are observed for Italy only in the second evaluation exercise, which had a stronger impact on the incentives for both researchers and universities.

Keywords Research assessment, performance-based university research Funding, university research, Italy

*The views expressed in this paper are only the authors' and do not reflect the official positions of the institutions which they are affiliated with. The usual disclaimers apply.

Massimiliano Bratti massimiliano.bratti@unimi.it
Tindaro Cicero tindaro.cicero@anvur.it
Enrico Lippo Enrico.lippo@unimi.it
Carmela Anna Nappi carmelaanna.nappi@anvur.org
Matteo Turri matteo.turri@unimi.it

[a] European Commission Joint Research Centre; Università degli Studi di Milano; IZA (Bonn)
[b] ANVUR
[c] Università degli Studi di Milano - Corresponding author. Dipartimento di Economia, Management Metodi Quantitativi (DEMM), Università degli Studi di Milano, via Conservatorio, 7, 20123 Milan (Italy). Telephone: +39 02503 21189.

1 Introduction

A growing number of countries have implemented over time national systems to evaluate the research outputs of universities and steer public funding to Higher Education Institutions (HEIs). According to Hicks (2012) at least 15 Performance-based University Research Funding Systems (PRFSs) were in place in 2012. Hicks (2012) defines a PRFS as a national exercise evaluating ex-post research outputs, and the consequent result-based fund allocation. The diffusion of PRFSs has spurred a heated debate on the way they are implemented and their impact on research in given disciplines. A number of studies have compared the different characteristics of these systems across countries (Genua and Martin, 2003; OECD, 2010; Northcott and Linacre, 2010; Hicks, 2012 and Rebora and Turri, 2013). A less developed stream of research revolves around the role that these evaluation systems play in the steering mechanisms between the Government and universities to determine whether the expectations underlying their implementation are met.

What benefits does a government (or a country) obtain, or expects to obtain, from a national research evaluation system? To answer this question many aspects of PRFs could be studied. Just to take a few, the significance of PRFSs in terms of legitimation (de Lange et al., 2010; Rebora and Turri, 2013), the opportunistic behaviours resulting from these exercises (Talib and Steele, 2000; Talib, 2003; Otley, 2010; Parker, 2011; Berlemann and Haucap, 2015: Edwards and Roy 2017), the effects on each discipline. This article focuses on the effect of PRFS on research, in particular on the output’s quantity/quality of each HEI. (The term Research Evaluation Exercise --- REE hereafter --- is used in what follows with the same meaning as PRFS, as it denotes a national exercise to evaluate ex-post research outputs whose results are linked to university funding.)

Although the impact of evaluation exercises has been discussed in many articles (see the following section), the literature often lacks data and sound empirical analyses proving their impact, and simply discusses country-specific case studies (Gläser, 2008; Butler, 2010). Robust evidence proving the link between the *introduction* of PRFSs and the increase of the quantity and quality of research output is very sparse. If, generally, there are some positive effects following the introduction of these exercises, the performance of HEIs in terms of research output is affected by the combination of multiple factors including the competition for non-governmental funds. Rigorous country-specific studies are thus required to explore the impact of PRFSs on each country (Auranen and Nieminen, 2010).

In early 2017, the Italian National Agency for the Evaluation of the University System and Research (ANVUR) published the results of the third National Research Evaluation Exercise. This was the last of three REEs, which were introduced since the early 2000s. In this paper, we leverage the Italian experience in research evaluation, one of the broadest in terms of scale and consolidated over 10 years (Rebora and Turri, 2013; Genua and Piolatto, 2016), to study how REEs impacted on scientific production in the last fifteen years.

We contribute to the existing literature in a number of ways. First, we focus on the Italian case study, which has not been extensively investigated yet, even though it has a significant dimension and history in terms of PRFS. Unlike the UK, where PRFSs were first implemented, Italy is a country characterised by a traditional university government system based on centralised regulation through laws, weak autonomy of universities and strong autonomy of the academic staff (Capano, 2003). Thus, this paper offers a better understanding of the potential benefits of PRFSs for governments by examining their implementation in a very different context from that existing in the country in which they were first introduced, namely the UK. The Italian case is also interesting since it offers variation over time in the amount of funding that was linked to the REEs' results and in the amount of research staff covered by the REEs, and allows to examine the effect of these specific incentivising mechanisms. A second contribution of this paper is mainly methodological. Indeed, we assess the impact of PRFS on research output using a *Difference-in-Differences in Reverse* (DIDR) strategy, by comparing a "switched" country, i.e. Italy, where a PRFS was firstly introduced in 2003, with the UK, a country which was always subjected to PRFSs in the period investigated in this paper. This strategy enables us to control for scientific area- and country-specific trends that might affect research quantity and quality in the two different countries, and is generally an improvement over the before-after comparisons existing in the literature, which may be affected by time-variant confounders.

The main results of this paper can be summarised as follows. First, we show that REEs were increasingly used by Italian governments to differentiate the allocation of funds to Higher Education Institutions (HEIs), progressively shifting from historical to merit-based funding criteria. The share of funds attributed to HEIs on the basis of REEs increased from less than 2% to 17% in 2016. Second, our analysis does not reveal any significant positive effect neither on the quantity of research output nor on research excellence, measured in terms of number of articles and percentage of top cited articles in Web of Science (WoS) database, respectively. On the other hand, our analysis demonstrates that PRFSs were associated with an improvement in the "average quality" of research output, measured by the percentage of cited articles in WoS. Yet, these effects were not generalised, but are only observed in the PRFSs which had an extensive coverage of the research staff, and linked a non-negligible amount of resources to theresults of the REEs (i.e. the second REE). We conclude that the introduction of PRFS is not sufficient per se to improve the average research output of a country, but its impact crucially depends on its ability to change the incentives of individuals and HEIs, and to increase their reactivity.

The paper proceeds as follows. The next section reviews the main goals related to the introduction of PRFSs and briefly summarises the extant literature on their effects. Section 3 describes ten years of PFRS experience in Italy, and the main changes that were introduced to the three REEs that came in succession in the last decade. Section 4 describes the empirical strategy and the data used to estimate the effect of REEs on the quantity and quality of research output. Section 5 comments on the main results and discusses some potential weaknesses of our analysis. The last section discusses our main findings and concludes.

2. Why introducing PRFSs?

Through national interventions with differences in both the timing and the solutions adopted but with similarities in their rationale, national universities systems in Europe have been reformed within a relationship between governments and universities that can be described as *steering-at-a-distance mode* (Paradeise et al., 2009; Huisman, 2009; Shattock, 2014). By playing a leading role at a distance, governments mainly employ PRFS mechanisms as a policy tool upon which the new template of systemic governance is based (Capano and Regini, 2014). The competitive performance-based university funding (Gornitzka et al. 2005; Lazzaretti and Tavoletti, 2006; Maassen and Olsen, 2007; Trakman, 2008; Capano, 2011) is linked to the reform of systemic governance in Higher Education and the need to justify, decrease or streamline public spending during a period of crisis (Teixeira and Koryakina, 2016).

The introduction of performance-based funding mechanisms is also closely related to New Public Management (NPM) ideology, which means that, in order to enhance competition in public organisations, there is a transition from an input-based funding system to an output-based funding system that rewards the providers (i.e. universities) that use the resources stemming from taxpayers in a cost-effective way (Weisbrod et al., 2010). It is no coincidence that the British RAE appears in many publications as one of the most prominent examples of NPM-inspired interventions (Power, 1997), so much so that Ferlie and Andresani (2009, p.187) define it as "the major NPM style instrument used to steer the UK academic field".

When, in the middle of 1980s, the British government decided to launch what will be the first national Research Assessment Exercise (RAE), the purpose was to allocate funding in a fair but non-egalitarian way (Ferlie and Andresani, 2009). As public spending was under review, the policy goal was to focus resources on the institutions that produced the highest-level research (Martin and Whitley, 2010). Through this measure, Margaret Thatcher's government intended to show the value for money of public funding allocated to British university research. To summarise, the spread of PRFSs is related with a reform process that has grown quickly in the last thirty years, i.e. the transition to a *steering-at-a-distance* governance model.

Another expectation is associated with the government's intention to stimulate through PRFSs research productivity of scholars in terms of the output quantity and quality.

Accounting researchers have shown for long the potential of performance measurement tools to produce specific behaviours in the evaluated subjects (Broadbent, 2010; Broadbent and Laughlin, 2009; Ferreira and Otley, 2009). Based on this assumption "national research assessment exercises can be analysed and managed as a management tool for promoting research performance in the way desired by central government bodies" (Rebora and Turri, 2013, p.1662).

In spite of this expectation, there is an ongoing debate about the link between the implementation of PRFSs and the increased quality and quantity of research produced at national level. The study of this topic is often based on single country studies given the local nature of PRFSs and the existence of very different country-specific contexts (Auranen and Nieminen, 2010).

The implementation of a PRFS in Australia was studied in relation to the increased productivity of the academic staff of Australian universities and the analysis showed that there was a positive outcome in terms of both productivity and increased publications on the most frequently-quoted journals (Butler, 2003). The adoption of a PRFS in Spain was shown to be related to the increased number of publications on journals in WoS (Jimenez-Contreras et al., 2003). Also Norway provides evidence in relation to the introduction of a PRFS in 2004 resulting in an increase in the scientific output of domestic academic staff (Butler, 2010). The case of the Czech Republic does not provide clear support for a positive relationship between the implementation of the PRFS and an increased research performance (Fiala, 2013; Vanecek, 2013). Finally, going back to the British RAE, there is a wide consensus among various authors that it significantly led to an increased research quality by improving the reputation and the attractiveness of the British university system (Brinn et al., 2001; Geuna and Martin, 2003; Otley, 2010). In particular, Moed (2008) showed that the academic staff tended to modify the scope of their research based on the criteria and the guidelines of the national research evaluation exercises. In 2014 this expectation led to the decision of the British government to evaluate the non-academic "impact" of research. The new exercise called REF (Research Excellence Framework) aimed to promote a behavioural change in the academic staff so that they could also focus on the non-academic outcomes of their research.

The British case study also highlighted that national evaluation exercises influenced several publishing choices and behaviours such as selection of methodological paradigms, multiple authorship and fractionate publications (Henkel, 1999; McNay, 2007; Hopwood, 2008; Parker, 2008; Fagerberg et al., 2012; Rafols et al., 2012).

The Italian case study has been recently examined in relation to the effects of the PRFS. Cattaneo et al. (2016), whilst analysing the productivity (i.e. average number of publications per researcher) of Italian universities, found that the implementation of evaluation exercises produced positive effects on productivity. A bibliometric analysis of research performance in two consecutive periods (before and after the VTR 2001-2003) highlights the increased productivity of Italian scholars. The study also showed how the specific characteristics of universities (e.g., legitimacy) can influence these effects.

3. Ten years of research evaluation exercises in Italy

In the last ten years Italy has launched three broad research evaluation exercises. All the exercises are based on the examination of research output, with a focus on international excellence, according to scientific areas; each area has a specific panel that establishes the methodological guidelines, manages the evaluation and makes a final judgement resulting in a ranking list.

The first exercise was launched in 2003: CIVR (Committee for Evaluation of Research), established in the Ministry of Education, implemented the exercise called VTR (Three-Year Research Evaluation) to evaluate research in the period 2001-2003 (Reale, 2008; Franceschet and Costantini, 2011). In November 2011, 8 years later, ANVUR (which replaced CIVR) launched a new research evaluation exercise focusing on the period 2004-2010 and called VQR (Research Quality Evaluation) (Ancaiani et al., 2015; Genua and Piolatto, 2016). In July 2015 ANVUR launched the third evaluation exercise (also called VQR) regarding the period 2011-2014. The three exercises are hereinafter referred to as V1, V2 and V3 for brevity.

The three REEs are characterised by a methodology based on the evaluation of research outputs in relation to their international quality. The exercises are organised in scientific areas with panels (called GEVs, Evaluation Expert Groups) in charge of establishing the methodologies, managing the evaluation and making a final judgement resulting in a ranking list. Some aspects were modified over time, especially between the first evaluation (V1) and the following two exercises (V2 and V3).

The first difference concerns the increased impact of bibliometrics compared to peer review. V1 examined research output exclusively through peer-review, whereas in the following exercises the panels used either or both methodologies: bibliometric analysis and peer-reviewing through external experts appointed by the panels. In V2 the panels were free to choose their methodology. However, about half of the research outputs under assessment were required to be evaluated by peer review. In V3 the role of bibliometrics was further enhanced through the distinction between “bibliometric disciplines” (Hard Sciences, Medicine, Engineering) and “non-bibliometric disciplines” (Humanities, Social and Economic Sciences). In bibliometric disciplines, the panel’s evaluation of papers was based on the combination of information regarding the scientific impact of papers, measured by the number of citations, and the impact factor of the publication outlet where they appeared. When the algorithm showed that there was a mismatch between scientific impact and impact factor, a paper was peerreviewed. In non-bibliometric disciplines peer review was still crucial, even though the panels could also use bibliometrics (this is what the Economics Panel did, for instance). However, also in V3 about half of the research outputs under VQR assessment were required to be evaluated by peer review.

Another significant difference between V1 and the following exercises regards the amount of outputs to submit and the submission process. In V1 universities were required to submit one research output published in the three years under examination (2001-2003) for every four scholars. In V2 the number of research outputs that each university had to submit increased to three papers published in the period 2004-2010 for each scholar; and in V3 two papers per scholar published in the period 2011-2014 were required. In 2003, 55,542 scholars submitted 13,585 publications; in 2010, with 62,709 scholars the expected number of scientific output subjected to evaluation increased tenfold to 150,000; whereas in 2015, with 50,354 scholars, the expected number of outputs was 100,000.

During the three evaluation exercises the number of scientific panels also varied. In V1 there were 20 panels, 14 scientific areas and 6 special interdisciplinary scientific areas. In V2 the evaluation was divided into 14 scientific areas with a close correspondence between the evaluated scholar and the panel. In V3 there were 16 scientific areas because two of the previous 14 areas were further subdivided to make panels more homogeneous.[1] Once again the most striking difference is between V1 and the following exercises. The joint effect of a fixed number of research outputs to be submitted by each researcher and the non-overlapping scientific panels for V2 and V3 ensured a closer link between research outputs, researchers and departments, which could not be found in V1. Subsequently, universities were less free in the selection of publications, which was crucial in V1, because there was no correspondence between outputs and researchers and there was no obligation to submit outputs for each researcher or scientific area. The panel’s reference area depended on the scientific area of the output, not of the authors (in Italy researchers are chategorised by discipline for the purpose of recruitment and promotion), and the submission of outputs for each scientific area depended on universities.

[1] Furthermore, in V3 an additional panel (the 17th panel) was established for the evaluation of the activities linked to the “third mission” (such as patents and spin-offs); however, this evaluation was implemented only for informational purposes and did not impact on funding.

In V1 the evaluation of each individual output was not released, while in the following exercises each scholar received her/his own evaluation for each publication that s/he (co-)authored. However, this individual evaluation of scholars was not released to universities, which could only be informed of the overall evaluation of departments or research groups in the same disciplinary sub-area.

As shown above, there are marked differences between the first exercises and the following ones. It was not just about the transition of responsibilities from CIVR to ANVUR (which absorbed CIVR) from 2011 onwards. The rational changed as well. V1 was limited to an assessment based on the ability of universities to produce a certain number of research outputs deemed excellent. Performance was linked to the quality of scientific outputs and the ability of universities' governing bodies to select the highest quality outputs in any sector. In V2 and V3, as each researcher was required to submit outputs, the exercise was not only aimed at assessing the quality of the best scientific outputs, but also the contributions of each scholar in their own areas; thus, evaluation also concerned the quantity of research, or differently said researchers' productivity, because it involved the research potential of each scholar over a multi-annual period.[2]

Importantly, the relationship between the results of REEs and the funding system slightly changed over time. In V1 performance was marginally used for fund allocation, whereas in V2 and V3 the Governments explicitly decided to link the results of the evaluation exercise with public funding of universities, so much so that the Ministerial Decree establishing V2 and V3 declared the decision of the Government to allocate increasing public funding to universities based on research quality as a precondition for the evaluation exercises. This aim became stronger over time, as shown by the Italian laws providing for an increasing share in public funding of universities to be based on research evaluation (Legislative Decree 213/2009 and Law 98/2013).

Although there has been a significant spread of research evaluation mechanisms across countries, at least since the creation of ANVUR in 2011, a main feature of the Italian case has been the absence of explicit university policies and formalised guidelines from the government (Capano et al., 2016). When compared with

[2] In all of the three exercises the evaluation of outputs was combined with the use of other indexes, such as indexes related to resource attraction and internationalisation.

similar European evaluation agencies, ANVUR is indeed striking for its strong independence from the Ministry of Education (Capano and Turri, 2017).

3.1 PRFSs as a funding tool

The Government has used V1 results to allocate funds since 2006. A growing part of the *Fondo di Finanziamento Ordinario* (FFO, Ordinary Fund), i.e. the lump sum that the Government allocates to public universities, has been allocated on the basis of PRFS results over time (Table 1).

From 2006 to 2012, based on V1 results, research performance outputs were constantly used to allocate university funding, but in variable proportions from 0.20% to 2.37%. A significant aspect was the update of the available data on research quality. Indeed, in the period 2009-2012 there was little use of V1 results because they were deemed insufficiently updated, concerning the period 2001-2003.

Table 1. Use of PRFS for public funding of universities (FFO, *Fondo di Finanziamento Ordinario*).

Year	Percentage of FFO allocated on the reward basis (%)[a]	Share of FFO allocated based on REE (%)[b]	PRFS employed to allocate funds
2006	3.6 (Evaluation model)	1.2	V1
2007	0.6 (Evaluation model)	0.2	V1
2008	2.7 (Evaluation model)	0.8	V1
2009	7.2 (Reward-based share)	2.4	V1
2010	10.29 (Reward-based share)	2.03	V1
2011	12 (Reward-based share)	1.58	V1
2012	12.8 (Reward-based share)	1.69	V1
2013	13.4 (Reward-based share)	8.84	V2
2014	17.3 (Reward-based share)	15.57	V2
2015	20 (Reward-based share)	17	V2
2016	20 (Reward-based share)	17	V3

[a] Based on factors such as outcomes of research, teaching, and other indicators.

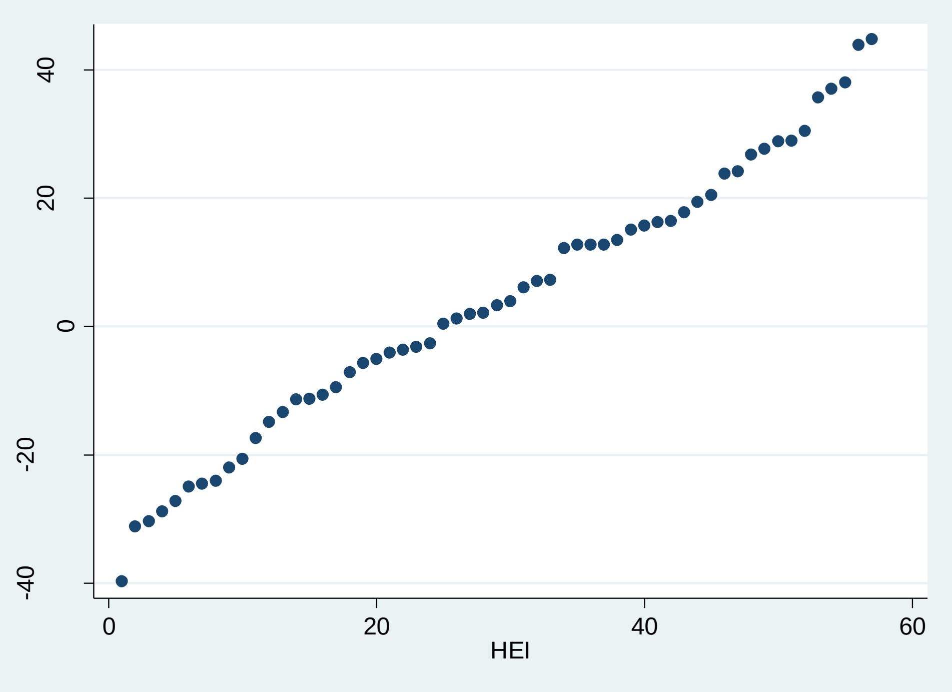
[b] It is a subset of the share allocated on a reward basis.

A major breakthrough took place with V2, whereupon increasing and large shares of FFO were allocated to universities. Since 2013 this link has become stronger, up to 17% of the available funding in 2015. Interestingly, in 2016 the distribution of the reward share to universities was postponed until the V3 results became available, i.e. by the end of the year.

Figure 1 shows the percentage difference between the share of total (national) merit-based funds allocated to each university according to PRFS (i.e. V2 results) in 2015 and the funds expected if the same total amount had to be distributed following the traditional history-based criterion. As can be seen, the differences can be remarkable, and as large as +40% or -30%.

Although the overall final allocation of FFO (i.e. the sum of the historical share, all the reward measures, and various supplementary funds) was affected by some mechanisms introduced to mitigate the effects of reward-based distributions (e.g. in 2015 no university was allowed to lose more than 2% of overall funding compared to the previous year), the implementation of PRFS led to a strong differentiation of funding especially because, as shown in Table 1, the share of reward-based FFO was increasing (Capano et al. 2016). Moreover, universities receive public funding from FFO to finance their costs, including personnel and administration costs, which are normally increasing over time. Albeit mitigated, the repeated reductions over the years of this funding required restrictive measures on the universities' budgets, impacting on the lives of institutions and scholars .
Besides, Law no. 98/2013 provides for a further increase in the share of reward-based funding, which will reach up to 30% of the overall FFO at full regime, and for four fifths of this funding to be based on REEs.

Fig. 1 Percentage gap of funding allocated to public universities based on PRFS compared to the funds that they would be assigned following historical attributions (in 2015)



Note. This picture shows the percentage variation between the amount of resources universities would have received if the merit-based quota were to be allocated based on historical criteria and those that they actually received according to the PRFS in force in 2015. The total amount of merit-based resources at the national level was 1.162 billion euros.

4 Methodology and data

4.1 Empirical strategy and choice of the "comparison" group

The main problem to be faced when one aims at evaluating the impact of national PRFSs on research output is the choice of a *comparison group*, which also represents the *counterfactual* that shows what would have happened in the absence of a PRFS. This is particularly difficult for the Italian case as all HEIs were subjected to the REEs exactly at the same time (i.e. there was no staggered implementation), and does not allow to exploit variation in the exposition across HEIs or geographic areas over time. Indeed, what have been done for Italy so far are *before-after* comparisons (e.g., Cattaneo et al., 2016). The idea behind this empirical strategy is to assume that had the PRFS not being in place, research output would have followed the same pre- PRFS trend. Thus, the pre- PRFS period provides the counterfactual. It is clear, however, that this counterfactual is obtained at a high cost: potential changes in research output attributable to time-varying unobservable variables (e.g. time trends), which cannot be controlled for in the empirical analysis, contribute to the quantification of the effect of the PRFSs. Key for the validity of this strategy is, therefore, that the model accounts for *all* potential determinants of research output, and that no relevant time-variant variable producing a time trend in the outcomes has been omitted. This assumption is rarely tenable, and we make an attempt in this paper to use what is generally considered as a more robust empirical strategy to estimate causal effects: a *Difference-in-Differences* (DD) type of estimator.

We choose the UK as the comparison group. The choice of the UK as the control group has been made considering the many similarities between the Italian and the UK's PRFS. Both PRFS have as the main object of evaluation the scientific production of academic researchers, in terms of publications: in both countries evaluation is carried out by panels of experts who are organised by research field, classify research outputs in merit classes and give them scores which in turn determine the final classification of the university. Finally, both PRFSs are probably the most extensive research evaluation exercises in Europe and this is why their comparison is meaningful as shown by Rebora and Turri (2013) and Genua and Piolatto (2016). The Italian assessment exercise has been, in fact, deliberately designed on the UK system (ANVUR, 2012), although there are some differences between these two systems in relation to methodology, administrative context and funding mechanisms.[3]

The similarity of the RAE, first, and the REF, later, in the UK with the Italian REEs, which were inspired by the former, makes the assumption that the PRFS implemented in the two countries can be considered as similar "treatments" more credible. We prefer this approach over: i) extending the analysis to a larger set of countries whose PRFS may have very different characteristics, making it hard to

[3] The main difference between the two exercises lays in that in UK researchers to be involved in the assessment program are selected by universities while the Italian VQR is a compulsory evaluation involving all staff with a permanent or temporary position in the universities and research institutes (Jappelli et al., 2015).

interpret the effects; ii) comparing Italy to countries where although national PRFS were never implemented, there might have been informal or partial (e.g. limited to some subject-groups) research assessment exercises affecting scientific productivity.[4]

The choice of the comparison group also dictates the choice of the empirical strategy. Indeed, in our empirical exercise we will be comparing a country, UK, which was *always treated* (i.e. where PRFS were always in place) to a country, Italy, which started to be treated after the implementation of the first REE. We define Italy as the "switched" group (in analogy to Kim and Lee, 2018). After the V1 both UK and Italy are treated. In this setting, Kim and Lee (2018) have recently shown that the causal effect of the treatment on the "switched" group during the pre-switch period can be estimated using a *Difference-in-differences in Reverse* (DDR) estimator.

The DDR approach has strong similarities with the traditional DD approach. Two main differences are however noteworthy. First, while the main identification assumption in the DD is the *common trend assumption* between the treated and the control units before treatment, the equivalent assumption in DDR is a common trend assumption after treatment. Second, based on these identification assumptions, while DD identifies the effect using differences in the treated and the untreated groups after treatment, DDR does it using differences between the treated and the "switched" groups before treatment.

DDR can be implemented parametrically using Ordinary Least Squares (OLS). In particular, let us define a, i, c, and t the scientific area, HEI, country and time subscripts (some of which will be used in later specifications), the following equation can be estimated using OLS:

$$Y_{iact} = \alpha_0 + \alpha_1 POST + \alpha_2 Q_c + \beta(Q_c * POST) + \epsilon_{iat}$$
(1)

where Y_{iact} is a measure of research output quantity or quality for scientific area a of institution i located in country c at time t; Q_c is a dichotomous indicator variable for the "switched" group, i.e. Italian HEIs; $POST$ is a dichotomous indicator for the post-switched period (i.e. after the introduction of the REEs in

[4] The German government, for instance, has run an excellence initiative awarding extra funding to the universities with the best future concept for research in 2006 and 2007.

Italy); ϵ_{iat} is an idiosyncratic error term.. The effect of interest is the coefficient on the interaction term between the “switched” group and the post-switched period, β.

The $POST$ and Q_c indicators can be replaced by year and country fixed effects (FEs, hereafter),D_t and D_c respectively, and the specification can be enriched by adding scientific-area FEs. Then equation (1) becomes:

MODEL 0 (area, country and year FEs)

$$Y_{iact} = \alpha_0 + D_a + D_c + D_t + \beta(Q_c * POST) + \epsilon_{iat} \; . \qquad (2)$$

Like DD, also DDR can be made more flexible to account for different trends for the treated and the “switched” units (Angrist and Pishke, 2009). In what follows, we introduce additional specifications which progressively saturate the model with time-trends and fixed effects:

MODEL 1 (area, country and time FEs; area- and country time trends)

$$Y_{iact} = \alpha_0 + D_a + D_c + D_t + \gamma_a * t + \gamma_c * t + \beta(Q_c * POST) + \epsilon_{iat} \; . \qquad (3)$$

MODEL 2 (area-country and time FEs; area-country time trends)

$$Y_{iact} = \alpha_0 + D_{ac} + D_t + \gamma_{ac} * t + \beta(Q_c * POST) + \epsilon_{iat} \; . \qquad (4)$$

MODEL 3 (area-institution and time FEs; area-institution time trends)

$$Y_{iact} = \alpha_0 + D_{ai} + D_t + \gamma_{ai} * t + \beta(Q_c * POST) + \epsilon_{iat} \; . \qquad (5)$$

where

- $\gamma_a * t$ is a scientific area-specific time trend;
- $\gamma_c * t$ is a country-specific time trend;

- D_{ac} is a scientific area-country indicator;
- $\gamma_{ac} * t$ is a scientific area-country-specific time trend;
- D_{ai} is a scientific area-institution indicator;
- $\gamma_{ai} * t$ is a scientific area-institution-specific time trend.

In the next section we estimate the effect of two treatments. One considering as the post-switched period the introduction of a PRFS in Italy, that is V1, and the other considering as the post-switched period the introduction of V2. Indeed, V2 marked a sharp change in the features of, and attitude towards, evaluation in Italy. First, because the V2 required all research staff and scientific areas to be evaluated (not only some selected research outputs or scientific areas on a voluntary basis like in V1). Second, as Table 1 shows, V2 also marked an important increase in the amount of funding of HEIs tied to their performance in the REEs.

Owing to the time lags existing between the first production and the final publication of scientific output, and the delays related to the inclusion in the bibliometric databases (namely WoS in this study), we consider the $POST$ period as starting with a delay with respect to the formal announcement of the REEs.[5] V1 was instituted with a Decree dated December 2003, accordingly we consider as the $POST$ period the one starting from 2006 (included) onwards. V2 was first announced in March 2010, but formally instituted with a Decree dated July 2011,[6] and we consider as the $POST$ period that going from 2013 (included) onwards. When evaluating the effect of V1, we consider the 2000-2012 period (i.e. omitting the potential effect of V2), while when evaluating V2 we focus on the 2001-2015 period, but in the spirit of a "donut-hole" regression we omit the 2006-2012 period which was affected by the V1 (and in which Italy cannot be considered as completely "untreated").

4.2 Data

[5] And we consider as the last period of the analysis 2015 to minimise the impact of the second type of delays for recent publications. WoS data were extracted in November 2016.

[6] This delay was due to the time needed to institute ANVUR which replaced CIVR, formerly in charge of the REE.

This analysis uses bibliometric data from Thomson Reuters Incites database which collects bibliometric data since 1981. We focus on a fifteen-year period ranging from 2000 to 2015. Data are organised by year, HEIs and scientific area. The sample used is composed of *academic* institutions that have taken part in the research assessments.[7] Each publication is associated to one or more HEIs on the basis of the authors' affiliations.[8] Publications are matched to six research areas defined by OECD: natural sciences; engineering and technology; medical and health sciences; agricultural sciences; social sciences; humanities.

Among the bibliometric indicators available in Thomson Reuters database, we have selected five indicators to be used as outcome variables: number of Web of Science documents; category normalised citation impact[9]; percentage of publications cited one or more times; percentage of publications in the top 10% of the citations' distribution; percentage of publications in the top 1% of the citations' distribution (by category, year and document type).

The first indicator allows us to explore the quantitative dimension of researchers' publication behaviour. An increase in the number of publications can be read as an increase in the scientific production of the researchers. The other indicators capture a qualitative aspect of publications. Citations can be generally interpreted as an impact indicator, as acknowledgment for the good quality of research from the scientific community and as a sign of the international circulation of the results. The last two indicators capture "excellence" measured by an article being placed in the top quantiles of the citation distribution. In this interpretation an increase in citations per paper or in percentage of papers cited would be understood as a positive effect of PRFS based on the international diffusion of research in the scientific community.

Figures A1-A5 in Appendix A show the scatter plots of the five research output indicators for the UK an Italy. Each point represents the mean of the indicator for a country's HEIs. Country-specific linear time trends for each period (pre-V1, V1

[7] The number of institutions that have participated in the research assessment exercises in Italy and are in the Incites database is 64. The number of UK academic institutions in the database is 124.

[8] In case of publications authored by researchers affiliated to different institutions, documents are computed in each institution. Hence the country's total does not coincide with the sum of publications in all institutions. In the Thomson Reuters database, publications can be associated to multiple research areas on the basis of the journal classification in those areas. Hence the sum of publications of all areas in a given country is higher than the total number of articles of the country.

[9] Number of citations per paper normalised for subject, year and document type.

and V2) are super-imposed to the scatter plots, and allow to check the post-treatment parallel trend assumption needed for the validity of the DDR. Figures A1, A2 and A3 show that the (post-switch) parallel trend assumptions seems to hold in the V1 period for the number of WoS documents and category normalised citation impact and the percentage of publications cited one or more times in WoS, while appears to be violated for the last two indicators. By contrast, for V2 the same assumption seems to hold for the percentage of publications cited one or more times in WoS, and to lesser extent for the percentage of WoS documents in the 10% and 1% of citations. Table 2 summarises this information.

Table 2. Validity of the post-switched parallel trend assumption

Research output indicator	**V1**	**V2**
n. documents	yes	no
category normalised citation impact	yes	no
% cited documents	yes	yes
% documents in top 10% of citations	no	yes
% documents in top 1% of citations	no	yes

Note. This table summarises the assessment of the validity of the post-switched parallel trend assumption, required for the DDR estimator, based on visual inspection of figures A1-A5 in Appendix A (as suggested by Kim and Lee, 2018). All indicators are computed on the WoS database.

5 Results of the Difference-in-Differences in Reverse (DDR) analysis

Table 3 shows the DDR estimates. Panel A reports the results for V1 and panel B for V2. The table shows only the DDR coefficients while the list of control variables (mainly FEs and time trends) is reported in the bottom part of the table. DDR coefficients can be roughly interpreted as the change produced by REEs in research output's quantity/quality with respect to a system without REEs during the pre-switched period, i.e. before 2006.

Focusing on the V1, a large but not statistically significant positive effect of about 17 publications per HEIs on the yearly number of WoS documents is estimated in Model 0. The magnitude of the coefficient is, however, greatly reduced and falls to 2.3 publications when area- and country-time trends are included in Model 1.

In the most saturated specification of Model 3 the coefficients becomes very close to zero (0.9). According to these estimates, Italy did not gain from V1 in terms of quantity of research output.

No statistically significant effect is found on the other outcome variables neither, although all coefficients are negative. Indeed, V1 does not appear to have impacted positively on the quality of the publications.

The picture emerging form Panel B is partially different. In this panel, the pre-V1 (2000-2005) and post-V2 (2013-2015) periods are compared. The positive effect estimated on the number of publications estimated with Model 0 turns into negative when country- and area-time trends are included in the empirical specification (Model 1). In Model 3 this negative effect is more precisely estimated and amounts to an about 38 publications reduction in WoS, statistically significant at 5%.

A statistically significant (at 5%) positive effect of 0.12 is instead estimated on the category normalised citation impact in the most parsimonious model (Model 0). However, also in this case the inclusion of time trends reduces the magnitude of the coefficient, which loses statistical significance. In the last column, Model 3 shows a coefficient very close to zero (0.11).

Quite interestingly, while in Model 0 the V2 appears to have negatively affected the percentage of the documents cited, when the outcomes are allowed to follow different country-, HEIs- and area-trends in Models 1-3, the effect turns into positive and is statistically significant at the 1%. The coefficients in Model 1, 2 and 3 are very stable, 9.4, 9.3 and 9.5, respectively. Thus, our analysis shows that V2 produced an increase in the percentage of WoS documents cited of slightly less than 10 percentage points. As we stressed before, this indicator mainly captures the "average quality" of research output.

By contrast, no effect is generally found on the two remaining outcome indicators, the number of documents on the top 10% and 1% of citations, respectively, which can be considered as proxies of "research excellence".

The estimates in Panel B exclude the 2006-2012 period since Italy was "partly" treated, i.e. subjected to V1. However, for the sake of completeness, we have also estimated the models for V2 on the whole 2000-2015 sample, considering the pre-2013 period as the pre-switched time span. The results (available upon request) are qualitatively consistent with those in Panel B, but the point estimates on the percentage of cited publications are smaller in magnitude. In particular, in the full sample, we find statistically significant positive effects both on the category normalised citation impact, namely 0.135 (s.e.=0.048), 0.147 (s.e.=0.073), 0.148 (s.e.=0.073) and 0.164 (s.e=0.078) in Models 0, 1, 2 and 3, respectively; and on the percentage of cited documents of 0.535 (s.e.=0.696), 2.948 (s.e.=0.919), 2.970 (s.e.=0.918) and 3.65 (s.e.=0.984) in Models 0, 1, 2 and 3 respectively, which are always statistically significant at the 1% level, except for Model 0.

Gathering together all results concerning V2, our analysis points to a possible reduction in the quantity of scientific production by Italian scholars (i.e. a reduction of the number of WoS articles), who have instead put more weight on the quality (i.e. impact) of their scientific output. In particular, a higher impact (i.e. a higher likelihood of being cited) could be achieved by scholars by engaging in more ambitious research projects and by targeting journals with a higher impact factor, in both cases with the consequence of improving their HEIs' performance in the REEs. However, according to our analysis such increase in average quality did not spur research "excellence".

The analysis in this section has confirmed that while V1 had no positive impact on the number of articles produced by Italian scholars, the quality of Italian research (as measured by the percentage of cited documents) improved thanks to the V2. How can the different results in the two panels be reconciled? Although the second result needs to be checked using a longer time period after the V2, our analysis suggests that the impact of PRFSs on the quality of publications crucially depends on the evaluation method adopted and the individual incentives it is able to create. As mentioned above, in V1 universities were required to submit only one research output for every four scholars, so the selection process involved only the publications of the best researchers working in each institution. This might have had a limited impact on the production of the "average" researcher, as he/she did not feel responsible for his/her institution's performance, which was only affected by top performers. Conversely, in V2 every scholar was required to submit multiple research outputs for the evaluation. Thus, all researchers became accountable for the good or bad performance of the HEIs they belonged to. Moreover, the individual results were disclosed to scholars in V2 but remained undisclosed in V1. Thus, whilst V1 was limited to an assessment based on the ability of universities to produce a certain number of research outputs deemed excellent, in V2 the evaluation was also aimed at assessing the research performance of every scholar, who had not received a feedback in the previous

REE. Based on these facts, and the larger amount of university funding linked to V2 (see Section 3.1), it is not surprising at all that an impact on average research quality is found after V2 but not after V1.

Table 2. DD4 results of the effect of VTR 2001-2003 (V1) and VQR 2004-2010 (V2)

Outcome	**Model 0**	**Model 1**	**Model 2**	**Model 3**
Panel A. V1 (VTR 2001-2003)				
N. documents	16.65	2.338	2.496	0.911
	(13.81)	(4.782)	(4.708)	(3.730)
Category normalised citation impact	-0.0253	-0.0843	-0.0848	-0.0784
	(0.0546)	(0.0800)	(0.0800)	(0.0864)
% cited documents	-2.143***	-1.068	-1.099	-1.748
	(0.691)	(1.235)	(1.229)	(1.224)
% documents in top 10% citations	-0.400	-1.271	-1.259	-1.225
	(0.581)	(1.087)	(1.082)	(1.134)
% documents intop 1% citations	-0.340	-0.203	-0.197	-0.112
	(0.226)	(0.392)	(0.391)	(0.383)
Period	2000-2012			
N. observations	13,085			
Panel B. V2 (VQR 2004-2010)				
N. documents	36.94	-26.23	-23.79	-37.76**
	(26.67)	(16.54)	(16.49)	(17.78)
Category normalised citation impact	0.122**	0.0365	0.0372	0.0114
	(0.0601)	(0.226)	(0.226)	(0.250)
% cited documents	-0.633	9.353***	9.343***	9.475***
	(0.838)	(3.115)	(3.110)	(3.176)
% documents in top 10% citations	0.498	3.398	3.509	3.535
	(0.661)	(2.844)	(2.844)	(3.058)
% documents in top 1% citations	-0.0990	-0.320	-0.294	0.145
	(0.277)	(1.249)	(1.251)	(1.256)
Period	2000-2005 and 2013-2015			
N. observations	9,033			
Control variables				

(scientific) area FE	Yes	Yes	No	No
country FE	Yes	Yes	No	No
year FE	Yes	Yes	Yes	Yes
area time trend	No	Yes	No	No
country time trend	No	Yes	No	No
area - country FE	No	No	Yes	No
area - country time trend	No	No	Yes	No
area - institution FE	No	No	No	Yes
area - institution time trend	No	No	No	Yes

Note. The Table reports the DDR coefficients ($Q_c * POST$) see Equations (1)-(5)) estimated with OLS on the dependent variables listed in the first column. The control variables are reported in the bottom section of the table. Standard errors are clustered by HEIs to allow for serial correlation in research output. *, **, *** statistically significant at 10%, 5% and 1% level, respectively.

5.1 Potential confounding factors

This section includes a brief discussion of other relevant changes in the institutional setting in the UK and Italy which might have impacted onthe research productivity of HEIs in both countries.

A potential confounding factor for our analysis is the institution of the National Scientific Habilitation (ASN) in Italy. With article 16 of Law 240 of 2010 the possession of the ASN became a necessary requirement for the participation in the universities' local competitions for the recruitment of Full and Associate Professors. The ASN is a non-comparative evaluation procedure managed directly by the Ministry of Education through the National Commissions by scientific area.

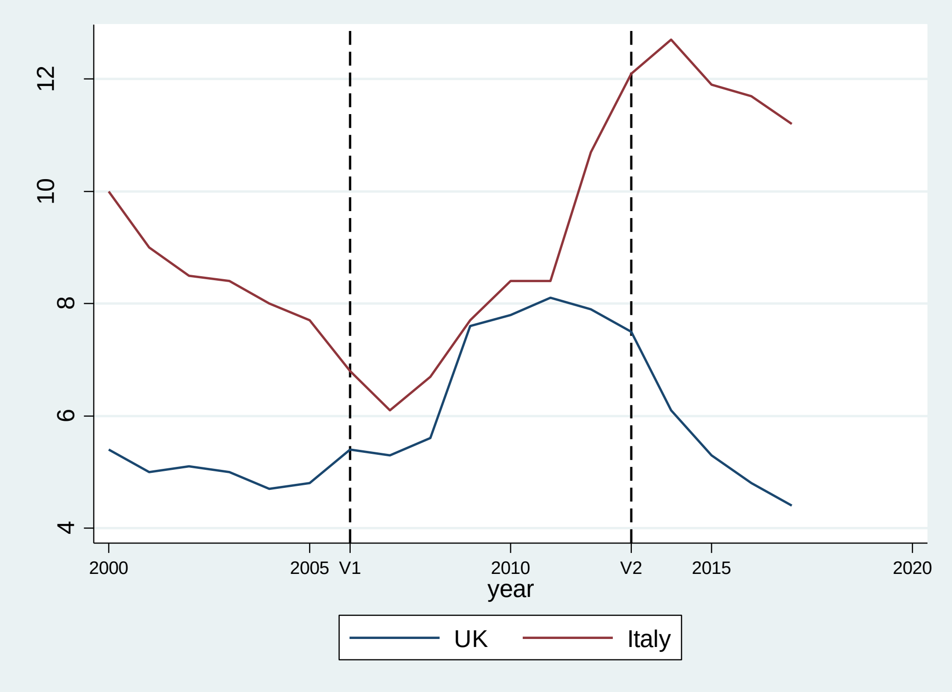
The ASN constitutes the qualification required to participate: 1) in local competitions organised by the universities with open (i.e. non-reserved) procedures (Article 18, Law 240/2010); 2) in competitions (up to 2019) reserved to those individuals who already have the qualification of Associate Professor for internal promotion to Full Professor (Article 24, paragraph 6, Law 240/10); 3) in recruitment procedures reserved to researchers of type B (tenure-track), who at the end of their three-year appointment can be hired as Associate professors (article 24, paragraph 5, Law 240/2010).[10] The first round of the ASN took place in 2012.

Since both the V2 and the ASN were created by the same law (Law 240/2010, the so called "Gelmini" Law, from the name of the Minister of Education), it is not possible to disentangle the separate effects of the two changes to the Italian higher education system. However, two considerations are in order. On the one hand, although the level of selectivity of the ASN was somehow different across scientific areas (ANVUR, 2016), the ASN may have introduced in the Italian system an additional meritocratic element for hiring and promotion, reinforcing the incentives to improve scientific production already produced by the REEs. On the other hand, this additional effect is uncertain since the REEs, especially after V2, may have already created enough incentives for HEIs to weight more the quality of scientific production in the hiring and promotion procedures of academic staff. In the UK, for instance, although an ASN was never in place, thanks to the research assessment exercises the quality of scientific production became a key element for hiring and promotion of researchers (Beattie and Goodacre, 2012; Oancea, 2014). In conclusion, the V2 may be considered as having created in Italy a system of incentives very close to the one that was already in place in the UK, making the ex-post parallel trend assumption needed for DDR credible.

A second potential confounder is the Great Recession, which, however, started to significantly impact the Italian and UK economies after the 2008, i.e. not at the same timing of the V1, and initially impacted the two economies in a similar fashion (see Figure 3). Hence, the recent economic crisis should not affect the estimates for the 2000-2011 period (i.e. our analysis of the impact of V1). However, the UK and Italian economies started to diverge, in terms of unemployment rates, from 2012. Although it is not easy to assess the effect of the economic crisis on scientific productivity, we think that the worse

Fig. 3. Unemployment rate in Italy and in the UK (2000-2017)

[10] Law 240/2010 created two types of temporary researchers (similar to Assistant Professors in the international context), type A, which is "non-tenure track" and type B, which is "tenure track". Type B researchers who obtain the ASN are hired as Associated Professor by the HEI for which they already work using special procedures described in article 24, paragraph 5 of Law 240/2010).



Note. Unemployment rate (number of unemployed/active population) for the UK and Italy (Source: our elaboration on Eurostat data).

performance of the Italian economy after the V2 may produce a downward bias in our estimates (i.e. bias them towards zero) making our estimates more conservative. On the other hand, Figure 2 shows remarkably parallel trends after V2 for some research output indicators, suggesting that the impact of the differential impact of the Great Recession, if any, might have been small, at least on some of the indicators analyzed in this study.

Among the events that have characterised the UK and, more specifically, England in the period under examination there is certainly the 2004 Higher Education Act, when HEIs could decide to charge fees up to £3,000 a year for undergraduate students and the subsequent provision, following the Browne Review in 2010, whereby the fee cup was raised up to £9,000 (Brown and Carasso, 2013). Although it represents a major change for British universities as it emphasises a stronger market-oriented approach, it is mainly concerned with teaching. There does not seem to be any impact on publications and thus on outputs.

The shift from RAE to REF in 2014 deeply affected research in the UK. The most important change was the inclusion of the non-academic impact of research in the evaluation with a percentage of 20% of the overall evaluation (REF, 2015). The transition to the REF, while representing a potential confounding factor, does not diminish the focus on publications, which is only accompanied by an extra factor, the non-academic impact: also in this case, the changes that have taken place do not affect the productivity of the HEIs.

6. Further discussion and concluding remarks

This paper investigates the Italian experience of adoption of PRFSs in the Italian case over fifteen years. The effects of two REEs (VTR 2001-2003 – V1- , and VQR 2004-2010 – V2-) are analysed.

Since 2016, following the V1 results and, even more significantly, since 2013 following the V2 results, increasing shares of government funding have been allocated to universities through PRFSs. For the Italian government, PRFS has become a pivotal tool to steer universities at a distance. Following NPM principles, universities have been increasingly funded on the basis of their actual research performance, thus reducing the operational funding allocated through more traditional criteria. This has impacted on the amount of resources received by each university compared to the amount received through the historically-based funding system.

The diachronic analysis of the use of PRFSs combined with the trend of the rate of state funding (FFO) allocated on a reward basis (not only based on REE but also on other indicators, such as the ability to attract research funding and teaching efficiency) suggests that the presence of PRFS affects the merit-based funding of universities but also that the implementation of this funding system is affected by other factors as well. In a highly regulated context such as Italy’s, funding is mainly influenced by laws providing for the adoption of criteria based on merit to allocate resources (Capano et. al, 2016). In this respect, the evidence collected not only highlights the importance of the political and institutional context to determine the success and the impact of a reform (Pollitt and Dan 2013; Bonini Baraldi, 2014), but also shows the potential of applying PRFSs to countries other than the UK, where this system was first experimented and whose research is more internationalised. The analysis of the data also suggests that since 2009 the enforcement of substantial cuts in public expenditure on universities has encouraged the use of reward-based allocation mechanisms (Capano et al., 2016). This evidence is in line with the experience of other countries, such as the UK (Teixera and Koryakina, 2016). In this respect, the presence of PRFSs is a technical solution to implement a reward-based allocation system of funding.

By using data on the number of papers published on journals included in the ISI Web of Science database along with some research impact indicators we intend to investigate whether REEs had any impact on the quantity and quality of research output. The analysis is carried out by comparing the trend of the publications of Italian universities and those of UK universities, using a *Difference-in-differences in Reverse* empirical strategy. The analysis has confirmed that while V1 had no positive impact on the quantity and quality of Italian scholars' research outputs, the "average quality" of Italian research as measured by the number of cited documents improved thanks to V2. Although this second result needs to be further investigated using a longer period after the V2, our results suggest that the impact of PRFSs on the quality and quantity of publications may crucially depend on the evaluation method adopted and the individual incentives it is able to create. In this respect, V2 unlike V1 involving all research staff of HEIs contributed to rising the accountability of each researcher who was now deemed responsible for the good or bad performance of his/her institution. The analysis points to the link between the effects of the evaluation on the scholars evaluated and the evaluation method adopted. Our paper shows that reactivity (Espeland and Sauder, 2007), i.e. the impact of assessment on the staff evaluated, is thus connected not to the introduction of PRFS mechanisms per-se, but to their structure and their detailed design.

A final consideration concerns the limitations of this study and the directions of future research. As shown in section 5, our analysis confirms that the PRFS in its recent configuration had a positive effect on the average quality of publications. This is useful to demonstrate in an evidence-based fashion that PRFSs brought some positive results; however, our analysis cannot explain how this occurred. In relation to the identified positive impact on research quality, what are the elements promoting reactivity from the academic staff? To what extent is this reactivity correlated with age of scholars, access procedures to the profession and career advancements? To what extent is it correlated with disciplinary or academic contexts favouring reactivity? The answer to these and related questions, which requires individual-level data, is left for future research.

References

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

Angrist, J. D., Pischke, J., (2009), Mostly Harmless Econometrics An Empiricist's Companion, Princeton: Princeton University Press.

ANVUR, (2012). Valutazione della qualità della ricerca 2004-2010 (VQR 2004-2010) Documento di accompagnamento dei criteri, http://www.anvur.it/wp-content/uploads/2012/02/documento_accompagnamento_criteri.pdf (accessed 12 February 2018)

Anvur 2016 RAPPORTO BIENNALE SULLO STATO DEL SISTEMA UNIVERSITARIO E DELLA RICERCA

Auranen, O., Nieminen, M., (2010). University research funding and publication performance – an international comparison. *Research Policy*, 39, 822–834. https://doi.org/10.1016/j.respol.2010.03.003.

Berlemann, M., Haucap, J. (2015). Which factors drive the decision to opt out of individual research rankings? An empirical study of academic resistance to change. *Research Policy*, 44, 1108–1115. https://doi.org/10.1016/j.respol.2014.12.002.

Beattie V., Goodacre A. (2012) Publication records of accounting and finance faculty promoted to professor: Evidence from the UK, *Accounting and Business Research*, 42, 197–231. https://doi.org/10.1080/00014788.2012.673159

Bonini Baraldi, S., (2014). Evaluating Results of Public Sector Reforms in Rechtsstaat Countries: The Role of Context and Processes in the Reform of the Italian and French Cultural Heritage System. *International Public Management Journal*, 17, 411-432. https://doi.org/10.1080/10967494.2014.935248.

Brinn, T., Jones, M.J., Pendlebury, M., (2001). The impact of Research Assessment Exercises on UK accounting and finance faculty. *The British Accounting Review*, 33, 333–355. https://doi.org/10.1006/bare.2001.0164.

Broadbent, J., (2010). The UK Research Assessment Exercise: performance measurement and resource Allocation. *Australian Accounting Review*, 20, 14–23. https://doi.org/10.1111/j.1835-2561.2010.00076.x.

Broadbent, J., Laughlin, R., (2009). Performance management systems: a conceptual model. *Management Accounting Research*, 20, 283–295. https://doi.org/10.1016/j.mar.2009.07.004.

Broucker B., De Wit K. (2015). New Public Management In Higher Education. In Huisman J., de Boer H., Dill D., Souto-Otero M. (eds.). The Palgrave International Handbook of Higher Education Policy and Governance. (pp. 57-75). New York: Palgrave MacMillan.

Brown and Carasso, 2013 Everything for sale? The marketisation of UK higher education

Butler, L. (2003). Explaining Australia's Increased Share of ISI Publications - The Effects of a Funding Formula Based on Publication Counts. *Research Policy*, 32, 143–155. https://doi.org/10.1016/S0048-7333(02)00007-0.

Butler, L., (2010). Impacts of performance-based research funding systems: a review of the concerns and the evidence. In: Performance-based Funding for Public Research in Tertiary Education Institutions: Workshop Proceedings, OECD Publishing, Paris, pp. 127–165, http://dx.doi.org/10.1787/9789264094611-en (accessed 29 November 2016).

Capano, G., (2003). Administrative Traditions and Policy Change: When Policy Paradigms Matter. The Case of Italian Administrative Reform During the 1990s. *Public administration*, 81, 781-801. https://doi.org/10.1111/j.0033-3298.2003.00371.x.

Capano, G., (2011). Government continues to do its job. A comparative study of governance shifts in the higher education sector. *Public Administration*, 89, 1622–1642. https://doi.org/10.1111/j.1467-9299.2011.01936.x.

Capano, G., Regini, M. (2014). Governance reforms and organizational dilemmas in European universities. *Comparative Education Review*, 58, 73–103. https://doi.org/10.1086/672949.

Capano, G., Turri, M. (2016). Same Governance Template but Different Agencies Types of Evaluation Agencies in Higher Education. Comparing England, France, and Italy. *Higher Education Policy*, 30(2), 225–243. https://doi.org/10.1057/s41307-016-0018-4.

Capano, G., Regini, M., Turri, M. (2016). Changing Governance in Universities. London: Palgrave-MacMillan.

Cattaneo, M., Meoli, M., Signori, A. (2016). Performance-based funding and university research productivity: the moderating effect of university legitimacy. *The Journal of Technology Transfer*, 41, 85–104. https://doi.org/10.1007/s10961-014-9379-2.

de Lange,P., O'Connell,B., Mathews,M.R., Sangster,A. (2010). The ERA: A Brave New World of Accountability for Australian University Accounting

Schools. *Australian Accounting Review*, 20, 24-37. http://dx.doi.org/10.1111/j.1835-2561.2010.00078.x.

Edwards, M. A., Roy, S. (2017). Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. *Environmental Engineering Science*, 34(1), 51–61. https://doi.org/10.1089/ees.2016.0223.

Espeland, W.N., Sauder, M., (2007). Rankings and reactivity: how public measures recreate social worlds. *American Journal of Sociology*, 113, 1–40. https://doi.org/10.1086/517897.

Fagerberg, J., Landström, H., Martin, B.R., (2012). Exploring the emerging knowledge base of "the knowledge society". *Research Policy*, 41, 1121–1131. https://doi.org/10.1016/j.respol.2012.03.007.

Ferlie, E., Andresani, G., (2009). United Kingdom from Bureau Professionalism to New Public Management. In Paradeise, C., Reale, E., Bleiklie, I., Ferlie, E. (Eds.), University Governance. Western European Comparative Perspectives. (pp. 177–196). Dordrecht: Springer.

Ferreira, A., Otley, D., (2009). The design and use of performance management systems: an extended framework for analysis. *Management Accounting Research*, 20, 263–282. https://doi.org/10.1016/j.mar.2009.07.003.

Fiala, D., (2013). Science Evaluation in the Czech Republic: The Case of Universities. *Societies*, 3, 266–279. https://doi.org/10.3390/soc3030266.

Franceschet, M., Costantini, A., (2011). The first Italian research assessment exercise: a bibliometric perspective. *Journal of Informetrics*, 5, 275–291. https://doi.org/10.1016/j.joi.2010.12.002.

Geuna, A., Martin, B., (2003). University research evaluation and funding: an international comparison. *Minerva*, 41, 277–304. https://doi.org/10.1023/B:MINE.0000005155.70870.bd.

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, 260–271. https://doi.org/10.1016/j.respol.2015.09.004.

Gläser, J., (2008). The social orders of research evaluation systems. In Whitley, E., Gläser, J. (Eds.), The Changing Governance of the Sciences. The Advent of Research Evaluation Systems. Sociology of the Sciences Yearbook. (pp. 245-266). Dordrecht: Springer.

Gornitzka, Å. M. Kogan and A. Amaral (eds.) (2005). Reform and Change in Higher Education – Analysing Policy Implementation. Dordrecht: Springer.

Henkel, M., (1999). The modernisation of research evaluation: the case of the UK. *Higher Education*, 38, 105–122. https://doi.org/10.1023/A:1003799013939.

Hicks, D., (2012). Performance-based university research funding systems. *Research Policy*, 41, 251–261. https://doi.org/10.1016/j.respol.2011.09.007.

Hopwood, A.G., (2008). Changing pressures on the research process: on trying to research in an age when curiosity is not enough. *European Accounting Review*, 17, 87–96. https://doi.org/10.1080/09638180701819998.

Huisman, J. (ed.). (2009). International perspectives on the governance of higher education. Alternative frameworks for coordination. New York: Routledge.

Kim, K., Lee, M. (2018). Difference in differences in reverse. Empirical Economics, published online: 05 June 2018. https://doi.org/10.1007/s00181-018-1465-0

Jappelli, T., Nappi, C. A., Torrini, R. (2015). Research Quality and Gender Gap in Research Assessment, *CSEF Working Papers*, 418, Centre for Studies in Economics and Finance (CSEF), University of Naples, Italy.

Jiménez-Contreras, E., de Moya Anegón F., López-Cózar D.E. (2003). The evolution of research activity in Spain: The impact of the National Commission for the Evaluation of Research Activity (CNEAI). *Research Policy*, 32, 123-142. https://doi.org/10.1016/S0048-7333(02)00008-2.

Lazzaretti, L., & Tavoletti, E. (2006). Governance shifts in higher education: a cross national comparison. *European Educational Research Journal*, 5(1), 18–37. https://doi.org/10.2304/eerj.2006.5.1.18.

Maassen, P., & Olsen, J. (2007). University dynamics and European integration. Dordrecht: Springer.

Martin, B.R., Whitley, R., (2010). The UK Research Assessment Exercise: a case of regulatory capture? In: Whitley, R., Gläser, J., Engwall, L. (Eds.), Reconfiguring Knowledge Production: Changing Authority Relationships in the Sciences and Their Consequences for Intellectual Innovation (pp. 51–80). Oxford: Oxford University Press.

McNay, I., (2007). Research assessment: researcher autonomy. In: Kayrooz, C., Akerlind, G., Tight, M. (Eds.), Autonomy in Social Science Research: The View from United Kingdom and Australian Universities. Elsevier, New York.

Moed, H.F., (2008). UK Research Assessment Exercises: informed judgments on research quality or quantity? *Scientometrics*, 74, 153–161. https://doi.org/10.1007/s11192-008-0108-1.

Northcott, D. and Linacre, S., (2010). Producing spaces for academic discourse: The impact of research assessment exercises and journal quality rankings. *Australian Accounting Review*, 20, 38-54. https://doi.org/10.1111/j.1835-2561.2010.00079.x.

Oancea Alis (2014) Research assessment as governance technology in the United Kingdom: findings from a survey of RAE 2008 impacts Zeitschrift für Erziehungswissenschaft November 2014, Volume 17, Supplement 6, pp 83–110 https://doi-org.pros.lib.unimi.it:2050/10.1007/s11618-014-0575-5

Organization for Economic Cooperation and Development (OECD), (2010). Performance-Based Funding for Public Research in Tertiary Education Institutions: Workshop Proceedings. OECD Publishing, Paris.

Otley, D., (2010). Research assessment in the UK: an overview of 1992–2008. *Australian Accounting Review,* 20, 3–13. https://doi.org/10.1111/j.1835-2561.2010.00074.x.

Paradeise, C., Reale, E., Bleiklie, I., Ferlie, E. (eds.). (2009). University governance. Western European comparative perspectives. Dordrecht: Springer.

Parker, J., (2008). Comparing research and teaching in university promotion criteria. *Higher Education Quarterly,* 62, 237–251. https://doi.org/10.1111/j.1468-2273.2008.00393.x.

Parker, L., (2011). University corporatisation: driving redefinition. *Critical Perspectives on Accounting,* 22, 434–450. http://dx.doi.org/10.1016/j.cpa.2010.11.002.

Politt, C. and G. Bouckaert. (2011). Public Management Reform: A Comparative Analysis New Public Management, Governance and the Neo-Weberian State, 3rd ed. Oxford: Oxford University Press.

Pollitt, C., Dan, S., (2013). Searching for impacts in performance-oriented management reform: A review of the European literature. *Public Performance & Management Review*, 37, 7-32. https://doi.org/10.2753/PMR1530-9576370101.

Power, M., (1997). The Audit Society: Rituals of Verification. Oxford: Oxford University Press.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., Stirling, A., (2012). How journal rankings can suppress interdisciplinary research: a comparison between Innovation Studies and Business & Management. *Research Policy*, 41, 1262–1282. https://doi.org/10.1016/j.respol.2012.03.015.

Reale, E., (2008). La Valutazione della Ricerca Pubblica. Milano: Franco Angeli.

Rebora, G., Turri, M., (2009). Governance in Higher Education: An Analysis of the Italian Experience», in J. Huisman (ed.), International Perspectives on the Governance of Higher Education. Alternative Frameworks for Coordination. New York: Routledge.

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

REF 2015 REF 2014 Manager's report http://www.ref.ac.uk/2014/pubs/refmanagersreport/

Shattock, M., (2014). International trends in university governance. London: Routledge.

Talib, D., (2003). Institutional behaviour impact of the 1996 RAE. *Higher Education Review*, 36, 57–77.

Talib, D., Steele, A., (2000). The Research Assessment Exercise: Strategies and Trade-Offs. *Higher Education Quarterly*, 54, 68–87. https://doi.org/10.1111/1468-2273.00145.

Teixeira, P., Koryakina, T., (2016). Political Instability, Austerity and Wishful Thinking: Analysing Stakeholders' Perceptions of Higher Education's Funding Reforms in Portugal. *European Journal of Education*, 51, 126-139. https://doi.org/10.1111/ejed.12126.

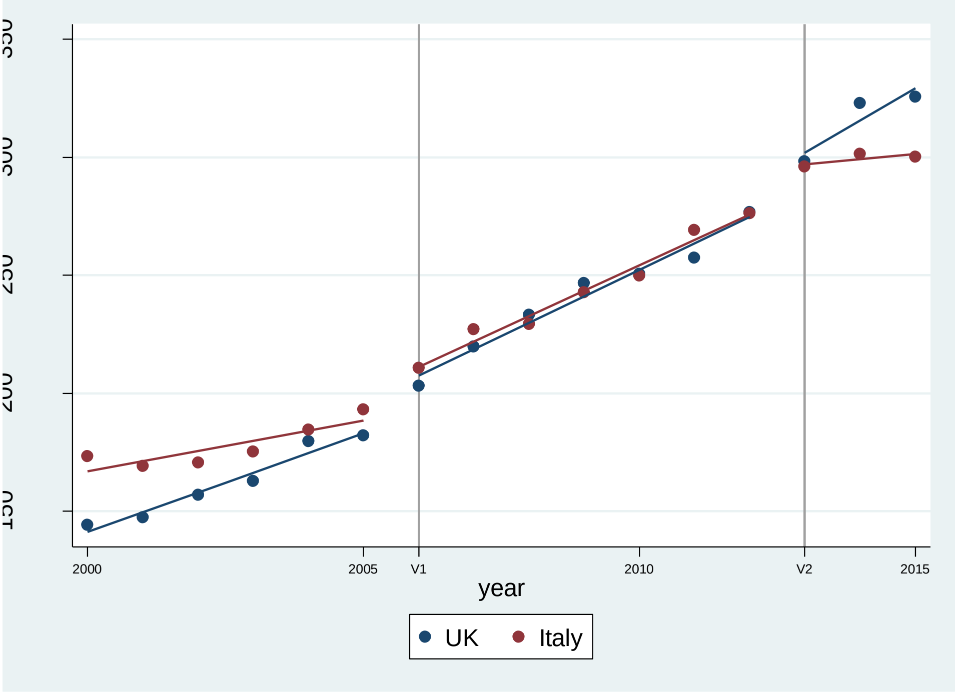
Trakman, L., (2008). Modelling university governance. *Higher Education Quarterly*, 62, 63–83. https://doi.org/10.1111/j.1468-2273.2008.00384.x.

Vanecek, J., (2014). The Effect of Performance-based Research Funding on Output of R&D Results in the Czech Republic. *Scientometrics*, 98, 657–681. https://doi.org/10.1007/s11192-013-1061-1.

Weisbrod, B.A., Ballou, J., Asch, D., (2010). Mission and Money: Understanding the University. Cambridge: Cambridge University Press.

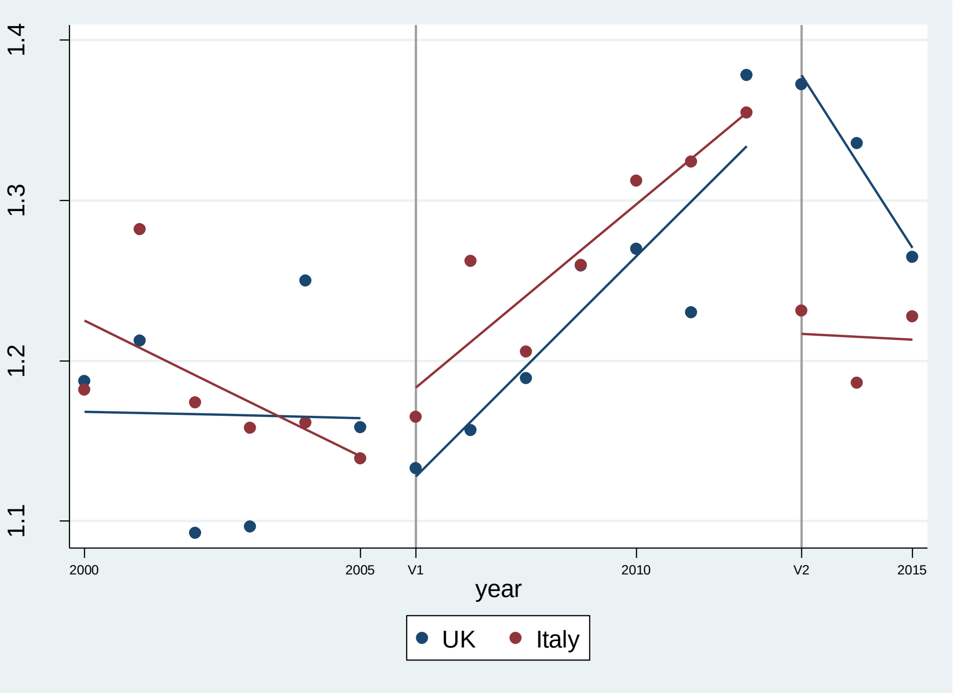
Appendix A. Check of the post-switched parallel trend assumption

Fig. A1. Trend of number of WoS documents



Note. Country-specific time trends by sub-period (pre-V1, V1, V2) are superimposed to the scatter plots.

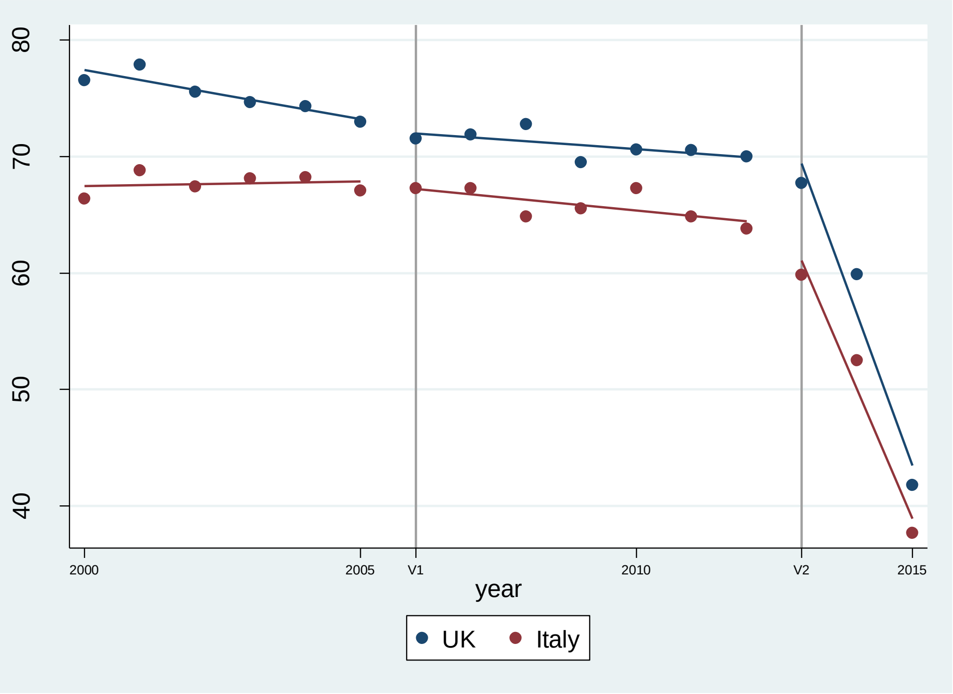
Fig. A2. Trend of category normalised citation impact



Note. Country-specific time trends by sub-period (pre-V1, V1, V2) are superimposed to the scatter plots.

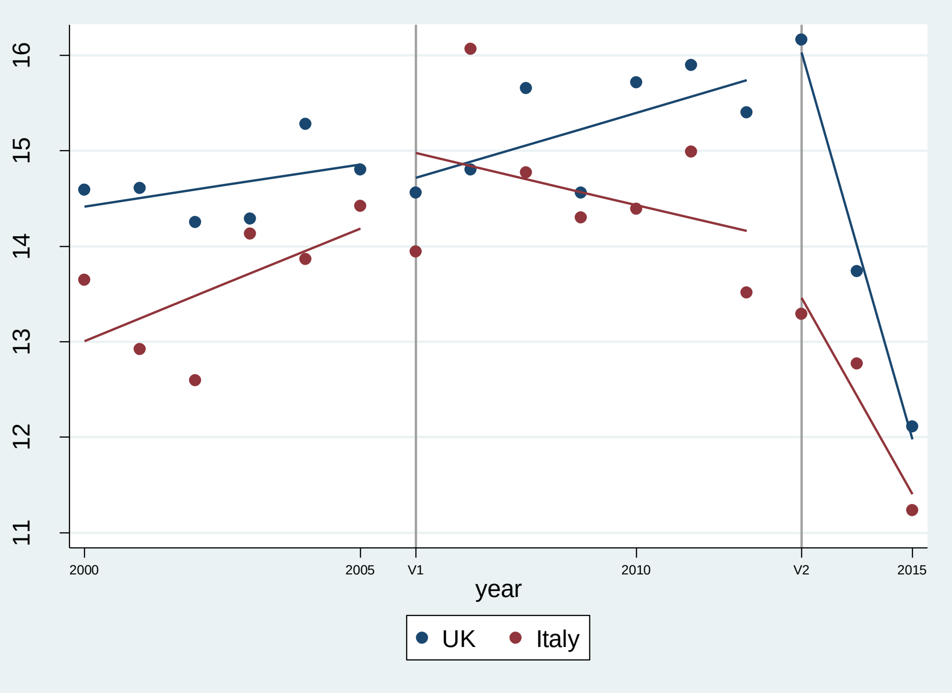
4

Fig. A3. Trend of percentage of publications cited one or more times in WoS



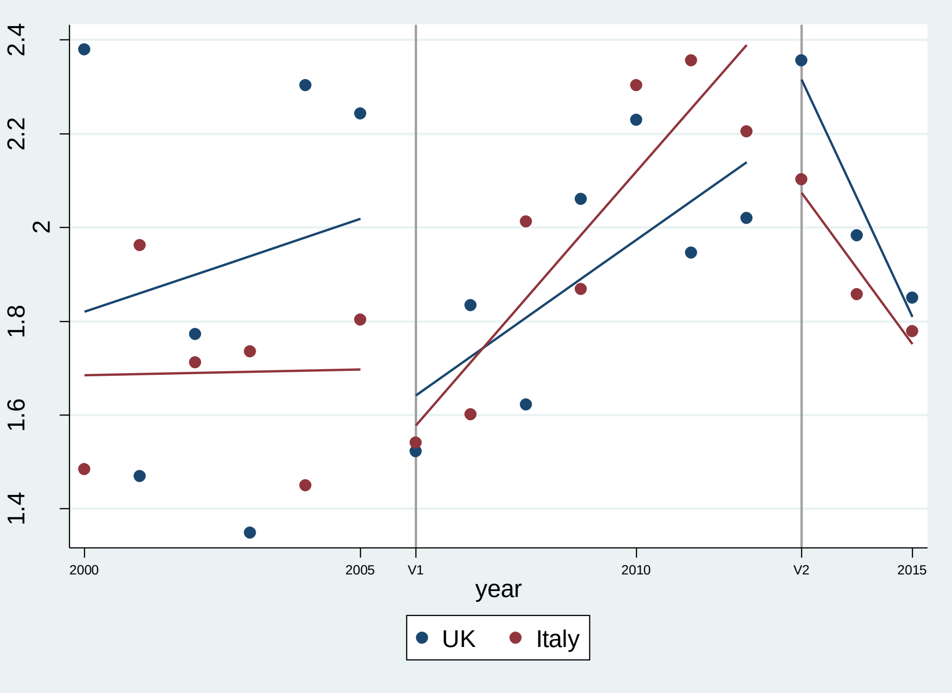
Note. Country-specific time trends by sub-period (pre-V1, V1, V2) are superimposed to the scatter plots.

Fig. A4. Trend of percentage of publications in top 10% of citations in WoS



Note. Country-specific time trends by sub-period (pre-V1, V1, V2) are superimposed to the scatter plots.

Fig. A5. Trend of percentage of publications in top 1% of citations in WoS



Note. Country-specific time trends by sub-period (pre-V1, V1, V2) are super-imposed to the scatter plots.