In Caperchione E., Bianchi C. (eds) *Governance and Performance Management in Public Universities*. SIDREA Series in Accounting and Business Administration. Springer, Cham. (pp. 49-68)

AN ANALYSIS OF METHODOLOGIES, INCENTIVES AND EFFECTS OF PERFORMANCE EVALUATION IN HIGHER EDUCATION: THE ENGLISH EXPERIENCE[1].

Giovanni Barbato[2], Matteo Turri[3]

Abstract

The chapter illustrates the English systems for research evaluation (the Research Excellence Framework - REF) and teaching evaluation (Teaching Excellence and Student Outcomes Framework – TEF) as NPM-driven mechanisms aimed at increasing public accountability and efficiency. Particular attention is given to the methodology of assessment, the incentives provided through the evaluation mechanism and the actual/potential effects that have been generating on both academics and universities. These two experiences are analyzed through the lens of the Management Control Theory (MCT), which argues that each evaluation system must be designed according to the nature of the activity/task under evaluation in order to be effective. Two dimensions are particularly relevant to this regard: the degree of the measurability and attributability of the output (i) and the knowledge of the cause-effect relation or transformation process producing the output (ii). Different configurations of these two dimensions lead indeed to different types of evaluation (input, process, output evaluation). Regarding the REF, it is highlighted how peer-review reflects both a process-based and an output-based evaluation since the quality of the research outputs under assessment should be judged according to the consistency of theoretical arguments and research methodologies employed to develop that specific research output. However, when this process shifts towards increasing attention to just quantifiable indicators of research outputs, some unintended consequences may arise. Concerning the TEF, the dominant focus of the metrics-based system towards measurable dimensions of teaching outputs, may not be able to grasp more qualitative and not directly measurable aspects that are still crucial in discriminating between satisfactory or partial performances.

Keywords: Research evaluation; Teaching evaluation, Research Excellence Framework; Teaching Excellence Framework; Management Control Theory

Introduction

The spread of performance evaluation (PE) systems in the higher education (HE) sector has been mainly driven by the New Public Management (NPM) narrative, as has occurred in other public sector domains (Ferlie et al., 2008). The NPM narrative assumes that the introduction itself of business-like processes, values and practices improves the efficiency, effectiveness and value-for-money of public sector activities. Regarding the HE sector NPM-inspired reforms have been concentrated in five main areas: a) introduction of competitive mechanisms for resources (students, funds) and performance-based funding; b) hardening of budgetary constraints; c) stress on performance evaluation (PE) and rankings; d) steering of HE systems through agencies and targets; e) strengthening of internal executive bodies and introduction of new managerial roles. PE is usually characterized by two main elements (Diefenbach, 2009): first, PE is a narrow conceptualization of performance in favor of quantifiable and short-term outputs measured through quantitative indicators; and second, PE is often linked to incentives that aim to orient and influence the behaviors of those who are evaluated in order to promote only specific behaviors and performances. In this sense, national PE systems of research and teaching can also be interpreted and analyzed as management tools used for promoting research and teaching performance in the way desired by central government bodies (Rebora and Turri, 2013).
Regarding the influence of NPM in the HE sector, the English reality is often seen as a clear example of a situation in which market-based mechanisms and PE mechanisms radically shape the operation of the HE

[1] This study is part of the PRIN project 'Comparing Governance Regime Changes in Higher Education: systemic performances, national policy dynamics, and institutional responses. A multidisciplinary and mixed methods analysis'.
[2] Dipartimento di Economia, Management e Metodi Quantitativi (DEMM), Università degli Studi di Milano: giovanni.barbato@unimi.it.
[3] Dipartimento di Economia, Management e Metodi Quantitativi (DEMM), Università degli Studi di Milano: matteo.turri@unimi.it.

system. The first national evaluation exercise of research can be traced back to the late 1980s, whereas teaching has recently seen the introduction of an annual assessment exercise since 2016 but with formers attempts during the '80s.
The purpose of this chapter is twofold. First, it aims to describe the rationales, features and effects of the evaluation of research and teaching in England. Secondary sources such as technical reports, official statistics and other empirical evidence from the literature have been used to carry out this goal.
Second, it analyzes both research and teaching evaluation through the theoretical lens of Management Control Theory (MCT), which specifically focuses on the relationship between the type of evaluation to be used and the features of the object to be evaluated (Ouchi, 1979; Frey et al., 2013). This perspective can be particularly useful to understand why NPM-inspired PE systems often generate unintended consequences or dysfunctionalities (Bevan and Hood, 2006; Diefenbach, 2009). While this theoretical approach has been fruitfully and widely applied to other public sectors (Frey et al., 2013; Barbato and Turri, 2017), it has been less applied to HE activities (Rebora and Turri, 2013). The chapter is organized as follows: the next section describes the MCT approach towards the evaluation system. Sections 3 and 4 illustrate the PE systems for research and teaching in England, whereas the last section discusses them through the MCT lens.

2. A theoretical framework for performance evaluation systems in universities

Management Control Theory (MCT) claims that the primary goal of performance measurement is to align individuals' efforts and organizations' objectives to improve performance. However, in order to effectively do that, the control system (here intended as the evaluation system) must be designed according to the nature of the activity/task under evaluation. In particular, two criteria must be considered (Ouchi, 1979; Frey et al., 2013), namely, the degree of the measurability and attributability of the output (high vs low) and the knowledge of the cause-effect relation or transformation process producing the output (high vs low).
The matching of these two criteria leads to three different types of evaluation/control, namely output control, input control and process control. The first typology focuses predominantly on the assessment of outputs given that the knowledge of the cause-effect relation producing the output cannot be known adequately and deeply. This type of evaluation is feasible and effective only when outputs can be easily measurable, in other words they can be observed and quantified. Moreover, outputs cannot be characterized by an intense interdependency between the actors that are involved in the generation of the output. In other words, output must be clearly attributable to individuals.
In contrast, when both output measurability and knowledge of the transformation process are low, the evaluation should instead be conducted on the inputs of the activity/tasks to be evaluated. In other words, it should verify that individuals in charge of that activity/task present specific knowledge and competences crucial for the generation of high-quality outputs (Ouchi, 1979). As described by Frey et al. (2013, p. 958), the input or 'clan' control assesses whether individuals 'have internalized norms and professional standards, i.e., are dedicated intrinsically to their task'. Key mechanisms that can be used to evaluate are thus the selection and career procedures and all the processes that socialize and train individuals towards the set of rules and code of practices required to carry out that specific task. In contrast to an ex-post output evaluation, input control is a long-term and more demanding process and is subject to a risk of low accountability towards those who do not share the same specific knowledge and rules.
The last typology provided by the MCT is process control, which assesses the processes carried out to achieve certain outputs. This can indeed be used when their measurability is not high, but evaluators display sufficient knowledge and shared understanding of the transformation process to generate them. Process control is characterized by a high transparency and equity of treatment. In contrast, it is argued that process control might fall into a bureaucratic and rigid procedure.
While MCT sheds light on the relevance of the relationship between the characteristics of the activity to be evaluated and the type of evaluation mechanisms that can be used, several scholars have widely underlined how NPM-based PE systems are predominantly skewed towards the assessment of measurable outputs. However, public services (e.g., education, health, security) are predominantly complex and unstable activities, with tasks highly interdependent so that outputs cannot be precisely attributed to individuals (Frey et al., 2013). The tendency of measuring only what can be quantified through indicators (named 'tunnel vision' by Smith 1995) entails that qualitative aspects are mainly neglected even though they might be crucial in defining whether individuals/organizations have actually achieved the desired output. This situation might even lead to

what Van Thiel and Leeuw (2002) describe as a 'performance paradox', whereby the evaluative metrics somehow misrepresent the actual level of performance achieved, since they lose their capacity to discriminate between satisfactory or poor performance.
Moreover, the provision of incentives may generate further unintended consequences if the type of evaluation adopted is not aligned with the features of the object to be evaluated. As claimed by Speklé and Verbeeten (2014), if output measurability is low and the knowledge of the cause-effect is ambiguous, the incentives will probably induce individuals/organizations to concentrate their efforts only towards the activities that are considered in the final evaluation or even on those more easily achievable despite their actual relevance (the so-called 'effort substitution'). Similarly, the concept of 'gaming' represents the voluntary manipulation/alteration of the evaluation process, which improves future evaluative judgment without substantially affecting performance. Several examples of 'gaming practices' have been documented across different public sectors (Van Thiel and Leeuw, 2002; Bevan and Hood, 2006; Diefenbach, 2009; Speklé and Verbeeten, 2014).
Therefore, thanks to its specific focus on the relationship between the type of evaluation and the characteristics of the activities/tasks to be assessed, the MCT represents a fruitful theoretical lens through which to analyze the PE system in the HE sector, especially in contexts where the NPM narrative played such an important role, such as in the English context.

3. The evaluation of research in England

Research at English universities is evaluated through a cyclical exercise, currently known as the Research Excellence Framework[4] (REF). The first evaluation exercise was introduced in 1986, followed by 1989, 1992, 1996, 2001, 2008 and 2014, with universities currently being involved in REF 2021. The evaluation exercise is managed by Research England[5], the current funding and evaluation agency for the English HE sector, on behalf of the other British nations. The REF introduction can be explained through a mix of financial and accountability rationales, with the predominance of the former (Martin 2011; Hicks 2012). Research activities were indeed previously funded according to antecedent historical allocations and based on student numbers. It was assumed that all academics were engaged equally in research activities, and thus universities could be funded almost identically despite their actual level of research orientation (Shattock, 2012). This situation relevantly changed with the public budgetary cuts undertaken by Thatcher's (1979-1990) governments as a result of the '70s economic crisis. The HE sector was also heavily affected. Consequently, the funding body of UK universities at that time, the University Grant Committee (UGC), decided to introduce an element of selectivity in its research formula funding by launching an evaluation exercise in 1986 to inform the distribution of increasingly shrinking resources (Technopolis, 2018). In this way, resources would have been allocated based on research quality, meeting the efficiency principle as well.
Finally, the introduction of an evaluation exercise also followed a rationale of public accountability; in other words, it was a way to verify that public expenditure was allocated effectively, making universities more accountable (Geuna and Martin, 2003; Bence and Oppenheim, 2005).

3.1 Main features of the Research Excellence Framework (REF)

The research evaluation system has increased in terms of research outputs evaluated and academics and universities involved over time. REF 2014 assessed 4 research outputs per academics (more than 191,000 outputs from 52,061 academics) and 154 universities against the 2 outputs per department from approximately

[4] The evaluation exercise has taken different denominations over time. Between 1986 and 1989 it was the Research Selectivity Exercise (RSE) while from 1992 to 2008 the Research Assessment Exercise (RAE). The evaluation exercise has then been renamed Research Excellence Framework (REF) from 2014. We will only use the term REF in this chapter to generally indicate the research evaluation exercise in England.
[5] Research England replaced the Higher Education Funding Council for England (HEFCE) through the 2017 Higher Education and Research Act. The funding agency was previously called the University Funding Council (UFC) between 1989 and 1992 and University Grant Committee (UGC) before 1989.

50 institutions in 1986. However, its features have mainly remained constant. This section describes them primarily on the basis of the last REF 2014.

I) *The REF is a peer-review evaluation system*, which means that the assessment regarding the quality of research is carried out by experts, in this case, other academics. The REF is one of the few research evaluation systems in the world that presents (almost) exclusively a peer-review methodology, whereas many of them employ a combination of bibliometric indexes and other output indicators (Hicks, 2012). REF evaluators are organized in disciplinary panels who are in charge of assessing the quality of the outputs submitted in the corresponding Unit of Assessment (UoA). The REF 2014 involved more than 1000 experts, 77% of whom were academics (890) appointed by the funding council. Reviewers are organized in 36 subpanels coordinated by 4 main panels. As the highest quality ratings awarded during the evaluation process refer to international excellence, each main panel has also involved some non-UK experts since 2001 to provide an exogenous viewpoint in calibrating the assessment criteria globally (Bence and Oppenheim, 2005). Moreover, panels also consist of research users (23%), namely, individuals coming from the business, culture and nonprofit sectors that either put universities' research into practice or benefit from it.

II) *The REF is organized in discipline-based units of assessment*. Each subpanel of evaluators corresponds to a UoA, which represents the unit of analysis of the assessment procedure. UoA roughly corresponds to a university department (Geuna and Martin, 2003). This discipline-based structure allows subpanels to specify evaluation criteria according to each discipline. However, to avoid losing consistency across disciplines, emerged during the first exercises, a limited number of 'coordinating panels' have been introduced since 2001 (Technopolis, 2018). These panels consist of the chairs of the subpanels to which they refer, and their main function is to improve the consistency of judgment criteria across them (Bence and Oppenheim, 2005). Moreover, these panels also contributed to addressing the (still) overlooked problem of interdisciplinarity.

III) *REF assessment is expressed in the form of ratings without aiming to build any ranking*. Ratings are indeed attributed not to the level of the entire university but to each UoA. However, the way ratings were awarded to the UoAs significantly changed with the 2008 exercise. A single-point rating was indeed granted for each UoA until 2001. In contrast, the 2008 exercise presented the results as quality profiles and not single-point ratings. A quality profile displays the percentage of the outputs submitted in each of the five judgment grades. Consequently, it sheds light on how quality is spread within each UoA, previously hidden by an overall single rating. This change had the effect of enlarging the number of UoAs that received funds, since also small pockets of research quality within low performing UoAa were going to be funded (Technopolis, 2018).

3.2 The object of assessment: research output, impact and environment

The REF has historically evaluated research by looking at its outputs. The definition of research output has been traditionally broad, including articles, books, book chapters, conference proceedings and patents (Geuna and Martin, 2003; HEFCE, 2015). The focus of the evaluation system was enriched with REF 2014, with the introduction of the evaluation of the nonacademic impact of research (Rebora and Turri, 2013). This choice was driven by financial and accountability reasons as well. Its introduction protected research funds from the Cameron government's (2010-2016) plan of budget cuts and was also a way to demonstrate the value of public investment (Martin 2011; Penfield et al. 2014).

As a result of several initiatives, including an impact pilot exercise (Technopolis, 2010), the funding agency ultimately included the research impact assessment in the upcoming REF 2014. Its weight in the final judgment is equal to 20%, with the quality of research outputs and research environment accounting for, respectively, 65% and 15% of the overall score. The impact was defined as 'an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia' (HEFCE, 2011, p. 48), and it should have been evaluated in terms of 'reach' (the spread or breadth of influence or effect on the relevant constituencies) and 'significance' (the intensity or the influence or effect).

Regarding the impact of research, universities must submit impact case studies, namely, documents that descriptively illustrate the impact and a template describing the piece of research (in the last 20 years) from which the impact has originated and their connection (HEFCE, 2015).

As shown by the impact pilot exercise, the methodology of case studies allows to adopt a broader view of the academic impact, especially beyond mere quantitative economic effects. Nevertheless, it also implies bureaucratic and time-consuming processes to develop them (Technopolis, 2010). Moreover, both Martin (2011) and Penfield et al. (2014) noted three main methodological challenges that could affect the overall goodness of the impact evaluation. First, the conceptualization and magnitude of the impact might be very different across disciplines in terms of the time lag between research and impact and in the degree of feasibility in gathering evidence. Second, not all the effects of academic research can be considered positive. Third, the impact might also be indirect and fuzzy and can evolve over time. This contrast, at least partially, is the linear model that underpins the funding agency's evaluation approach on the impact assessment. Last, Samuel and Derrick (2015) show that evaluators can present heterogeneous perceptions about how to characterize impact, pointing to a potential issue of subjectivity and inconsistency of judgement as well.
'Research environment' is the third dimension of research evaluated by the REF, and a template describing the research environment is submitted for each UoA. This is assessed in terms of its contribution to the 'vitality' and 'sustainability' of the unit under evaluation. The template consists of four sections: research strategy; academic and nonacademic staff; income and infrastructures/facilities; and contribution to the discipline.

3.3. Effect of research evaluation and transition towards REF 2021

This section highlights the main intended and unintended consequences that the evaluation exercise has generated at the level of the HE sector, the university and academics (Rijcke et al., 2016). Some of these have also been clearly identified by the last independent review of the REF chaired by Lord Nicholas Stern (Stern, 2016).

I) *At the level of the HE sector*, the REF has certainly contributed to enhancing the research orientation of British universities in terms of both productivity and awareness, especially for the more 'teaching-oriented' institutions (Bence and Oppenheim, 2005; Moed, 2008). However, the concentration of incentives on research has undeniably decreased the value and importance being given to teaching from both academics and universities (Henkel, 1999; Geuna and Martin, 2003; UCU, 2013), especially within the most research-intensive institutions.
Second, the REF-based funding system has increasingly concentrated resources within a limited set of institutions that receive the predominant share of quality-related funds, leading to a mechanism in which the 'rich get richer'. The 1992 exercise allocated approximately 90% of funds to pre-1992 universities, whereas approximately 75% of REF 2014 funding was concentrated in Russell group universities, with the top 10 universities receiving just over 50% of the funds. Furthermore, this concentration implies that it would be difficult for less research-oriented universities to invest more in research activities and improve their performances much further. Ultimately, this concentration is almost entirely self-perpetuating (Geuna and Martin 2003).

II) *At the level of universities*, the REF provides information as well as benchmarking possibilities that can be used for the internal strategic management of the university (Technopolis, 2015). Furthermore, REF results can also be employed to externally improve the university's image.
Nevertheless, the evaluation exercise has also generated increasingly significant costs for universities (Geuna and Piolatto, 2016). The main share of REF 2014 costs is indeed covered by universities, and it is estimated at £232 million out of the total £246 million (Technopolis, 2015). This amount is much higher than that of the 2008 (£66 million) and 1989 (£4.1 million) exercises, partly due to the introduction of the impact evaluation. The largest percentage of the costs is represented by submission preparation (£212 million), among which £55 million only concerns the impact (£7500 pound per impact case study). Moreover, the 'Accountability review' has also depicted an increasing bureaucratization of the evaluation system within universities, with the introduction of procedures and mechanisms aimed to support the REF submission process (Technopolis, 2015). These range from appointing a REF manager/committee to run a mock REF to choose the staff and outputs worth submitting. Most of these employ bibliometrics to predict the outputs that were going to be evaluated more positively. Several universities have also hired external consultants to develop impact case studies.

III) *At the level of academics*, three main effects can be identified. First, the REF affects the nature and character of the research carried out by academics (Butler, 2007; Technopolis, 2018). Adams and Gurney's analysis (2014) shows an increasing tendency from academics to skew the selection of their outputs to be submitted towards articles published in journals with a high impact factor, even though these articles are not well cited or even research papers. This is particularly evident in some fields of study in which articles are not the first primary source of publications, such as the social sciences and engineering. Being published in such journals is seen by itself as an objective and immediate proxy of research quality. Harley (2002) similarly sheds light on this process of compliance with the perceived demands of the evaluation exercise. Moreover, Joynson and Leyser (2015) and Technopolis (2018) illustrated a widespread perception among academics that interdisciplinary works could not be perceived as positively by university management as monodisciplinary works in terms of future REF evaluation (UCU, 2013). However, this issue stems more from how universities prepare submissions rather than from the mere presence of the REF (Technopolis, 2018).
Second, academics are increasingly subjected to psychological pressures to publish enough papers to be considered for the submission process. The 'publish or perish' narrative is indeed argued to lead to a commodification of academic labor (Henkel, 1999; Bence and Oppenheim, 2005). Almost 60% of the University and College Union's survey respondents (2013) recognize that the pressure to meet REF expectations had increased their stress levels, and approximately 35% also experienced negative health effects. This can further affect their research integrity, encouraging the manipulation of data or the reporting of only positive results, as reported by Joynson and Leyser (2015).
Third, the REF is claimed to affect academics' careers directly and indirectly (Technopolis, 2018). Some studies have revealed that academics are often sanctioned if their research performances were not considered satisfying, with some cases of early forced retirement (Bence and Oppenheim, 2005). In terms of recruiting, many universities have started to use bibliometrics to determine who is going to contribute most to improving future REF submissions (UCU 2013).

Three main changes in the next REF 2021 (Research England, 2019) can be relevant to address some of these issues. First, universities could previously choose which academics can submit, causing potential negative consequences on academics. REF 2021 establishes that universities will now have to return all staff with 'significant responsibility for research'. Universities still decide upon who displays this feature, but this has to be justified through the development of an internal code of practice concerning the selection of outputs. A second change addresses the number of outputs that can be submitted. REF 2021 shifts the emphasis from individuals to the submitting unit in computing the number of outputs to be submitted, partly reducing pressure to produce publications. Third, research outputs could no longer be 'portable' from one university to another as a result of the academics' transfer if the outputs were already made publicly available. Finally, REF 2021 also broadens the interpretation of impact, including that on public engagement and teaching; furthermore, it introduces an institutional level assessment for the research environment and provides new instruments to address interdisciplinary research.

4. The evaluation of teaching in the English HE sector

Similar to research activity, quality of university teaching is mainly[6] assessed through a national evaluation exercise known as the Teaching Excellence and Student Outcomes Framework (TEF).
The potential introduction of the TEF was initially mentioned in the Conservative party's manifesto for the 2015 general election. This identified its rationale in ensuring that universities deliver the best possible value for money to students, which significantly contributes to the HE sector through high tuition fees and the connected loan system. A second purpose of the TEF is to inform students' decisions about where to study based on the conception of students as 'consumers', which stems from the increasing marketization of the English HE sector (Gunn, 2018; Deem and Baird, 2019). However, the influence of the TEF in this regard seems still limited. A survey carried out by the Universities and Colleges Admissions Service (UCAS) (2018) highlights that only 17.1% of students approaching HE know what the TEF is, and 58% of this group stated that it was important in deciding which university to choose.

[6] The set of quality assurance processes implemented by the Quality Assurance Agency (QAA) operates alongside the TEF. Yet, this chapter will only focus on the TEF since this is the main evaluative system that counterbalance the REF.

The 'market' narrative attached to the TEF has drawn several critics towards it (Deem and Baird, 2019), which have been further fostered by its rapid and top-down implementation from the Department for Education (DfE) and the funding agency (Gunn, 2018). A first introductory year of the TEF was indeed carried out in 2016 after an open consultation about its metrics, and it was legally implemented through the 2017 Higher Education and Research Act.
Finally, the TEF was also thought to counterbalance the excessive emphasis on research brought by the REF by providing both reputational and financial incentives towards teaching, even though these incentives seem modest compared with those provided by the REF. Reputational incentives are expressed in the gold, silver or bronze medals awarded during the assessment procedure, while the financial incentives included the possibility (not introduced to date) to charge tuition fees up to £9250 per year instead of the current maximum cap of £9000. Nevertheless, one of the first pieces of empirical evidence (Cui et al., 2019, UUK, 2019) underlines that the TEF certainly contributes to rebalancing attention towards teaching, even though the majority of academics (and universities) still attribute a very small or no effect to teaching quality.

4.1 Features and operation of the Teaching Excellence Framework (TEF)

The TEF is a metrics-based evaluation system and thus evaluates teaching quality through a set of indicators covering different areas of the teaching mission, namely, 'teaching quality', 'learning environment' and 'student outcomes and learning gain' (DfE, 2017). The Office for Students[7] (OfS) is the body in charge of coordinating and managing the TEF. In contrast to the REF, it assesses teaching quality every academic year and provides a single rating at the institutional level. However, the government is aiming to implement the TEF at a subject level in the next several years due to two subject-level pilot exercises conducted for a sample of universities. The results of these pilot exercises will also inform Shirley Pearce's independent review of the TEF that is expected to provide recommendations in 2021, leading to a potential significant revision of the TEF evaluative framework.
The TEF evaluation procedure is carried out by an independent panel of experts and assessors, of which two-thirds are academics and one-third are students. Some experts on widening participation and employment are involved too. Moreover, TEF officers will provide training and assistance to panel members/assessors during the entire assessment process. The evaluation process is based two main sources of evidence: the quantitative indicators reported in Table 2 and a qualitative and narrative document known as 'provider submission'.
The quantitative indicators are divided into six core metrics and three supplementary metrics. The metrics related to 'teaching quality' (1a and 1b) and one of 'learning environment' (2a) are measures of students' satisfaction calculated on a set of questions from the National Student Survey[8] (NSS). The other indicators regard the status (3a) and type (3b) of employment of graduates as well as the regularity of students' careers (2b). As can be noted from Table 2, almost all the metrics are measures of outputs or outcomes (career progression; employability) of the learning and teaching activity while processes and inputs are less considered and only through the perspectives of students (e.g. feedbacks from teachers/tutors). The metrics are computed for the three most recent years, cover only the undergraduate provision, and are presented separately for full-time and part-time students (OfS, 2018). Furthermore, each core metric is computed for a series of subgroups based on some characteristics of the student body, such age, gender, ethnicity, disability and domicile. These are known as split metrics. All the metrics are calculated directly by the OfS through national databases on HE and managed through a TEF metric workbook. The workbook provides for each core and split metric a benchmark value and the difference between the university's metric value and the benchmark, along with a z-score that reports if the difference is statistically significant (this is underlined with a flag). The benchmark is a measure of 'expected performance', a weighted sector average for that specific metric, that takes into account variables that are outside the control of the provider, such as the entry qualification of students and the subject of study, to mention a few. Both the benchmark and the difference thus inform assessors on how to interpret metric values and are indeed unique for each university.

[7] As a result of the 2017 Higher Education and Research Act, the Office for Students has replaced HEFCE as the regulatory body for teaching.
[8] The NSS records students' opinions on several aspects of their degree programs in the final year of their academic career.

Table 2. TEF core and supplementary metrics, description and source of the data

Areas	Core and *supplementary* metrics	Description and source
Teaching quality (1)	Student satisfaction with teaching on their courses (1a)	National Student Survey[9] (NSS), questions 1 to 4 (at 2018): - Q1: staff are good at explaining things. - Q2: staff have made the subject interesting. - Q3: staff are enthusiastic about what they are teaching. - Q4: my course has challenged me to achieve my best work.
	Student satisfaction with assessment and feedback (1b)	National Student Survey (NSS), questions 8 to 11 (at 2018): - Q8: the marking criteria have been clear in advance. - Q9: Marking and assessment has been fair. - Q10: Feedback on my work has been timely. - Q11: have received helpful comments on my work.
	Grade inflation (1c)	HESA[10] and ILR[11] student records: This metric provides information on the type of degrees awarded in order to give evidence on the grading policy. It measures the number/proportion of the different types of degrees awarded (1sts, 2:1s, other degree classifications and unclassified degree awards) 10 years ago and in the most recent three years.
Learning Environment (2)	Student satisfaction with academic support (2a)	National Student Survey (NSS), questions 12 to 14 (at 2018): - Q12: I was able to contact a member of the staff when I needed to. - Q13: I received sufficient advice and guidance in relation to my course. - Q14: good advice was available when I needed to make study choices.
	Student retention on courses (2b)	HESA and ILR student records: This metric tracks students from the year they enter a HE provider to the next academic year. To be counted as continuing, a student must either have qualified or be recorded as actively studying on a HE programme.
Students outcomes (3)	Employment or further study (3a)	DLHE[12] survey: Percentage of UK-domiciled leavers who say they are working or further studying at 6 months after graduation.
	Highly skilled employment or further study (3b)	DLHE survey: Percentage of UK-domiciled leavers who say they are in highly skilled employment[13] or studying at 6 months after graduation.
	Above median earnings (3c)	LEO[14] dataset: Proportion of qualifiers in sustained employment who are earning over the median salary for 25 to 29 year-olds.
	Sustained employment or further study (3d)	LEO dataset: Proportion of qualifiers in sustained employment or further study three years after graduation.

Source: Office for Students (2018): TEF Framework. Year Four Procedural Guidance

[9] The NSS records students' opinions on their degree programs and on the overall academic support during the entire academic career by final year students on a voluntary basis.
[10] Higher Education Statistical Agency (HESA).
[11] Individualized Learner Record (ILR).
[12] Destinations of Leavers in Higher Education (DLHE).
[13] The UK Standard Occupational Classification (SOC) classifies jobs within 1, 2 and 3 groups as 'highly skilled'.
[14] Longitudinal Education Outcomes (LEO) dataset.

The second source of evidence used during the assessment process is the provider submission. This is a qualitative document, no longer than 15 pages, through which universities contextualize their own performances and illustrate their institutional approach towards teaching excellence and how this affects students (OfS, 2018). In this regard, as shown by Beech (2017), several additional qualitative and quantitative data are reported, such as citations from the external quality assurance review and student union statements, other internal learning analytics, UCAS data and other national league tables and rewards.
The assessment process is structured in three consecutive steps (DfE, 2017) and carried out by an independent panel of experts and assessors as aforementioned. During the first step, panel members will only look at the core metrics with attention to their distance from benchmarks (the flags) and using split metrics and contextual data when necessary. The three metrics build up on NSS data present a weight of 0.5 while the others equal to 1. Based on this process an initial hypothesis on the rating is generated[15]. Secondly, the provider submission and the supplementary metrics will be then considered to decide if the initial hypothesis can be confirmed or has to be modified (the second step). Both the first and the second assessment steps are operatively carried out within small groups of panel members that consist of (at least) two academics and (at least) one student. Each group will look at a set of HE providers. Finally (third step), a meeting of the full TEF panel will determine collectively the final rating (A 'Gold', 'Silver' or 'Bronze' medal) based on the recommendations advanced by each group of panel members. A statement of findings is also provided to each HEI, in which the reasons behind the award are explained. Although metrics are undeniably more important in the overall assessment process, Beech (2017) has highlighted that approximately 25% of the initial hypotheses has been changed after the analysis of the provider submission.

5. Discussion and conclusion

Previous sections have descriptively illustrated the main features of the PE of research and teaching activities in England, giving particular attention to their patterns of implementation (policy rationales), methodology of assessment and intended (and not) effects on academics and universities. This section aims to analyze the two aforementioned experiences by using the theoretical lens of the Management Control Theory (MCT),. and the two analytical dimensions introduced in section 2, namely (I) the degree of the measurability and attributability of the output and (II) the knowledge of the cause-effect relation producing the output (or transformation process). The evaluation of research will be discussed initially and subsequently that of teaching. In terms of evaluation methods, the assessment of research carried out in the REF, is based on a peer review process conducted by experts that assesses the quality of a research output (a publication, a patent) according to the consistency of theoretical arguments and research methodologies employed to achieve that specific research finding, in other words, according to knowledge of the transformation process that is behind the development of that research output (Frey et al., 2013). The strength of the peer review method is thus the reference by the evaluators (usually other academics) to the same code of practices, competences and knowledge that are embodied by the object of assessment (research products) (Turri, 2005). In this sense, the peer review method balances elements coming from both the process and output control types described by the MCT in section 2.
An evaluation process focusing uniquely on outputs might be indeed problematic since the attributability and measurability of research outputs can vary, especially for team-based disciplines (e.g. natural sciences and medicine), in which the specific contribution of each researcher to the output might not be completely transparent and thus identifiable. Furthermore, the capacity of the few quantitative metrics available on research outputs (e.g., number of citations, journal impact factor) to exclusively and comprehensively assess the quality of research is controversial and the object of a still-open debate (Weingart, 2005; Butler, 2007; Rijcke et al., 2016). In contrast, knowledge of the cause-effect process leading to a research output should always be high due to the replicability and transparency requirements of science, favoring a process-oriented evaluation like peer-review. Moreover, research outputs are 'singularities', in other words, unique and original, and cannot be assessed through standardized and mechanic methods (Rebora and Turri 2013).
Therefore, at least in theoretical terms, the REF methodology seems to be consistent with the nature of research and the values of the academic community.

[15] So, for example, if a HE provider presents a total value of 2.5 based on its core metrics, should be assessed, at the end of the first step, with the 'Gold' rating.

However, the literature has noted some dysfunctionalities related to the REF (Harley, 2002; Geuna and Martin, 2003; Rijcke et al., 2016). These unintended consequences are unsurprisingly the result of a 'silent' drift in the conceptualization of research evaluation towards a mere quantitative evaluation of outputs (the output control typology of MCT). An important driver of the increasing importance of metrics can be primarily found in how scholars perceive research quality, which has increasingly become skewed towards specific quantitatively measurable properties of research outputs (e.g. articles vs books; journal's impact factor). These are increasingly seen by academics (and university management) as objective metrics of good performance and thus used to select papers for the future REF submission. As underlined by Adams and Gurney's study (2014, p. 8), academics tend to 'favour journals with high average citation impact and among those journals they are persuaded that a high Impact Factor beats a convincing individual article'. Moreover, universities' top management started to internally implement their own set of evaluation practices, often relying on quantitative metrics as well as setting targets, incentives and sanctions that usually affect career progression (UCU; 2013). Therefore, an excessive relying on measurable aspects of research outputs from both academics and university top management, can hamper the fragile balance between the specific features of the activity under assessment and the appropriate evaluative instruments as underlined by the MCT. This process can also generate unintended consequences on the quality and authenticity of individual (and organizational) research performances such as manipulation of data and low-quality papers to meet deadlines, as highlighted by Joynson and Leyser (2015).

Second, as mentioned in the theoretical section, MCT emphasizes the scope of gaming behaviors that can arise from incentives when the typology of evaluation adopted does not fit well with the activity to be assessed. As underlined by Speklé and Verbeeten (2014, p. 132), if the quantitative metrics will only provide a partial representation of a task/activity, the presence of strong incentives might 'induce organizational actors to focus on target achievement rather than on organizational goals'. Regarding the REF, two main gaming practices can be identified from the literature. A first example is represented by the practice of universities hiring top researchers shortly before the REF census date to improve future REF ratings by increasing their research performance (Bence and Oppenheim, 2005; Stern, 2016; Technopolis, 2018). Besides highlighting a case of 'gaming' this example clearly shows how the evaluation system will also misrepresent the actual level of research quality, leading thus to a case of performance paradox (Van Thiel and Leeuw, 2002), whereby the evaluative mechanism loses its capacity to discriminate between satisfactory or poor performance due to the blurring effect generated by this 'gaming' behaviour.

A second example of performance paradox can be instead identified in the way universities develop the 'research environment template' (see section 3.2). Despite the standardized structure of this template, it has empirically shown how universities tend to misrepresent the actual research environment by dwelling on past achievements rather than current efforts and by using several writing 'ornaments' with an uncommon ratio of adjectives to verbs (Thorpe et al., 2017). As underlined by Wilsdon et al. (2015, p. 129), 'the narrative elements were hard to assess, with difficulties in separating quality in research environment from quality in writing about it'.

Third, while peer review seems to naturally fit with the features of the research activity, it also entails relevant and increasing costs. Geuna and Martin's (2003) analysis suggests that while initial benefits may outweigh the costs, a burdensome system such as the REF will probably generate decreasing returns over time, thus raising questions about its future sustainability and the efficiency of the peer-review methodology itself (Geuna and Piolatto, 2016). These unintended consequences have been partly softened and addressed by the evidence-based reviews after each evaluation exercise, with particular relevance of the last one (Stern review) illustrated at the end of section 3.3.

Regarding the evaluation of teaching activities carried out through the TEF, the analytical lens of the MCT allows to highlight a complex task whose assessment is anything but linear and effortless. Teaching activities are indeed characterized by a low and unclear attributability and measurability of outputs as well as by an unstable and unpredictable knowledge of the cause-effect relation producing the output. First,

teaching is characterized by the intensive interdependency between teacher and student efforts, which are both crucial for the success of this activity (Turri, 2005). The success of teaching is certainly shaped by teaching/pedagogical competences displayed by the teacher, but it equally depends on students' attitudes and efforts. These are also the result of prior educational paths and personal and intellectual capacities that are not necessarily generated within the teacher-student interaction. This feature implies not only a problem of attributability of outputs but also that the knowledge of the cause-effect relation is intrinsically low due to the instability, unpredictability and complexity of this activity.

Second, while research activity naturally tends to be synthesized in a product whose assessment gives a good approximation of the research process behind it, the teaching activity does not display a similar representative output, and several activities might be recognized as such (Turri, 2005). Some can be measured through indicators (e.g. graduates' employability or students' dropout), while others can be quantified less (e.g. students' learning gain) (Leiber, 2019). Based on what has been illustrated by the MCT, the features of teaching activity illustrated so far, would suggest that the mere evaluation of some quantitative outputs would not be enough to capture the complex nature of teaching. Nevertheless, the TEF is strongly oriented towards the assessment of a narrow set of outputs (Gunn 2018), covering only those that are more easily measurable through quantitative measures, namely, student satisfaction and graduates' employability. This entails that other more qualitative but still-crucial aspects of teaching, such as learning gains, teachers' competences/attitudes, learning strategies and curriculum design (Leiber, 2019; Cui et al., 2019), are omitted from the assessment process, resulting in the abovementioned issue of 'tunnel vision' (Smith, 1995).

Furthermore, teaching outputs that are quantified and assessed through the TEF (student satisfaction and graduates' employability) are claimed to be only limitedly shaped by the teaching process since other exogeneous factors intervene in their generation. Concerning the students' satisfaction metrics (Table 2, no. 1a, 1b, 2a), it is claimed how students' satisfaction and teaching quality are different constructs since satisfaction is influenced by factors that are beyond the teaching process itself (Spooren et al., 2013). Metrics regarding the employment of graduates (metrics 3a and 3b) are instead claimed to be partially affected by factors that do not depend on universities' efforts in teaching activities but are related, for example, to the health of the local labor market and economy (Deem and Baird, 2019; UUK, 2019). The ambiguity on which factors really and ultimately affect teaching outputs under assessment (such as satisfaction and employability), might thus weaken the ability of the evaluative metrics to discriminate between a good or a poor performance, resulting in a potential 'performance paradox' problem. Last, TEF metrics cover only undergraduate provision and mostly UK domiciled students, since international students are not included in employment metrics, although they represent approximately 20% of HE students in England. In summary, it is often claimed that the TEF measures teaching quality only indirectly, with questionable metrics and with a narrow perspective, resulting in a partial representation of teaching quality within universities (O'Leary and Wood, 2019; UUK, 2019).

Concerning the emergence of unintended consequences related to the evaluation of teaching activities, even though evidences on the effects of the TEF on academics and universities are still limited, some potential risks might already be envisaged.

First, since TEF metrics do not fully capture all relevant teaching dimensions but the evaluation exercise provides reputational incentives to universities (in the form of the 'gold', 'silver' or 'bronze' medal), university management might implement strategies and invest resources with the narrow goal of essentially improving and prioritizing the activities measured through metrics (e.g., employability), thus risking to lose and not supporting a more holistic vision of teaching. This aforementioned unintended consequence is known as 'effort substitution' (Speklé and Verbeeten, 2014) . The first empirical evidence on the TEF illustrates similar tendencies. Cui et al. (2019) illustrate indeed how the TEF has certainly increased the internal centralization and standardization of teaching activities as well as the accountability of academics, with activities mainly directed at satisfying TEF metrics and not improving the overall L&T experience.

Second, the high relevance of student satisfaction metrics could provide negative incentives for universities to actually lower teaching quality, since 'innovative forms of teaching […] often score low student satisfaction ratings, despite these methods often being highly effective in enhancing student learning' (RSS, 2016, p. 1). These 'gaming behaviors' would hardly be spotted by the metric nature of TEF assessment, potentially undermining its effectiveness.

Finally, MCT underlines that when the measurability/attributability of outputs and knowledge of the cause-effect process are low, input control, based on the selection and socialization of individuals towards that specific task/activity, can be more effective (Frey et al., 2013). Concerning the evaluation of teaching, this insight might be interpreted as a call to shift attention from exclusive output evaluation to the potential value of training young academics in teaching activities (instead of just focusing on research training and production) as well as faculty development.

In conclusion, the interpretation of these two experiences of PE through the lens of the MCT certainly contributes to shed light on the importance of the relationship between the nature/features of the activity/task under evaluation and the most appropriate instrument/methodology to be used for its assessment. This relationship can indeed be crucial to understand more deeply potential and actual unintended consequences

that PE systems can generate, as it has been illustrated in this chapter, and should thus properly be considered by both national and local policymakers during the design of PE systems.

Acknowledgements

We acknowledge the support by the Italian Ministry of Education, University, and Research through the PRIN 2015: 'Comparing Governance Regime Changes in Higher Education: systemic performances, national policy dynamics, and institutional responses. A multidisciplinary and mixed methods analysis' (2015RJARX7).

References

Adams, J., and Gurney, K. A. (2014). *Evidence for excellence: has the signal overtaken the substance? An analysis of journal articles submitted to RAE2008.* London: Digital Science.

Barbato, G., and Turri, M. (2017). Understanding public performance measurement through theoretical pluralism, *International Journal of Public Sector Management*, 30(1), 15-30.

Beech, D. (2017), *Going for Gold: Lessons from the TEF provider submissions*, Higher Education Policy Institute (HEPI) Report 99, Oxford: UK.

Bence, V., and Oppenheim, C. (2005). The evolution of the UK's research assessment exercise: publications, performance and perceptions. *Journal of Educational Administration and History*, *37*(2), 137-155.

Bevan, G. and Hood, C. (2006). What's measured is what matters: targets and gaming in the English public health care system, *Public Administration*, 84(3), 517-538.

Butler, L. (2007), Assessing University Research: A Plea for a Balanced Approach, *Science and Public Policy*, 34(8), 565–74

Cui, V., French, A. and O'Leary, M. (2019). A missed opportunity? How the UK's teaching excellence framework fails to capture the voice of university staff, *Studies in Higher Education*, DOI:10.1080/03075079.2019.1704721.

Deem, R. and Baird, J. (2019). The English Teaching Excellence (and Student Outcomes) Framework: Intelligent Accountability in Higher Education?. *Journal of Educational Change* (Article in press).

Department for Education (Dfe) (2017). *Teaching Excellence and Student Outcomes Framework Specification*. October 2017. London: DfE.

Diefenbach, T. (2009). New public management in public sector organizations: the dark sides of managerialistic enlightenment, *Public Administration*, 87(4), 892-909.

Ferlie, E., C. Musselin, and G. Andresani. (2008). The Steering of Higher Education Systems: A Public Management Perspective. *Higher Education*, 56(3), 325–48. doi:10.1007/s10734-008-9125-5.

Frey, B.S., Homberg, F. and Osterloh, M. (2013). Organizational control systems and pay-for-performance in the public sector, *Organizational Studies*, 34(7), 949-972.

Geuna, A., and Martin, B. R. (2003). University research evaluation and funding: An international comparison, *Minerva*, 41(4), 277-304.

Geuna, A., and Piolatto, M. (2016). Research assessment in the UK and Italy: Costly and difficult, but probably worth it (at least for a while), *Research Policy*, *45*(1), 260-271.

Gunn, A. S. (2018a). Metrics and methodologies for measuring teaching quality in higher education: Developing the teaching excellence framework (TEF), *Educational Review*, 70 (2), 129–148.

Harley, S. (2002) The Impact of Research Selectivity on Academic Work and Identity in UK Universities, *Studies in Higher Education*, 27(2), 187-205.

HEFCE (2011). *Research Excellence Framework 2014: Assessment framework and guidance on submissions, REF 02/2011*. Bristol: United Kingdom.

HEFCE (2015). *Research Excellence Framework 2014: Manager's report*. Bristol: United Kingdom.

Henkel, M. (1999). The Modernisation of Research Evaluation: The Case of the UK, *Higher Education*, *38*(1), 105-122.

Hicks, D. (2012). Performance-based university research funding systems, *Research policy*, *41*(2), 251-261.

Joynson, C., and Leyser, O. (2015). The culture of scientific research. *F1000Research*, *4*(66), 1-11

Leiber, T. (2019). A general theory of learning and teaching and a related comprehensive set of performance indicators for higher education institutions, *Quality in Higher Education*, 25(1), 76-97.

Martin, B. R. (2011). The Research Excellence Framework and the 'impact agenda': are we creating a Frankenstein monster?. *Research evaluation*, 20(3), 247-254.

Moed, H. F. (2008). UK Research Assessment Exercises: informed judgments on research quality or quantity?, *Scientometrics* 74 (1), 153-161

O'Leary, M. and Wood, P. (2019). Reimagining teaching excellence: why collaboration, rather than competition, holds the key to improving teaching and learning in higher education, *Educational Review*, 71(1), 122-139.

Office for Students (2018). *Teaching Excellence and Student Outcomes Framework (TEF) Framework. Year Four Procedural Guidance*. OfS 45/2018. Bristol: United Kingdom.

Ouchi, W. (1979). A conceptual framework for design of organizational control mechanism, *Management Science*, 25(9), 833-848

Penfield, T., Baker, M. J., Scoble, R., and Wykes, M. C. (2014). Assessment, evaluations, and definitions of research impact: A review. *Research evaluation*, *23*(1), 21-32.

Rebora, G., and Turri, M. (2013). The UK and Italian research assessment exercises face to face, *Research policy*, *42*(9), 1657-1666.

Research England (2019). *Guidance on submissions – REF 2021 01/2019*. Bristol: United Kingdom.

Rijcke, S. D., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use: a literature review, *Research Evaluation*, *25*(2), 161-169.

Royal Statistical Society (RSS) (2016). Response to the Department for business innovation and skills' technical consultation (year 2) on the teaching excellence framework. Available at: http://www.rss.org.uk/Images/PDF/influencing-change/2016/RSS-response-to-BIS-Technical-Consultation-on-Teaching-Excellence-Framework-year-2.pdf.

Samuel, G. N., and Derrick, G. E. (2015). Societal impact evaluation: Exploring evaluator perceptions of the characterization of impact under the REF2014. *Research Evaluation*, *24*(3), 229-241.

Shattock, M. L. (2012). *Making policy in British higher education 1945–2011*. Maidenhead, England: McGraw-Hill/Open University Press.

Smith, P., (1995). On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration* 18, 277–310.

Speklé, R.F. and Verbeeten, F.H.M. (2014). The use of performance measurement systems in the public sector: effects on performance, *Management Accounting Research*, 25(2), 131-146

Spooren, P., Brockx, B., and Mortelmans, D. (2013). On the validity of student evaluation of teaching, *Review of Educational Research*, 83 (4), 598–642.

Stern, N. (2016). *Building on success and learning from experience: an independent review of the Research Excellence Framework*. London: UK.

Technopolis (2010). *REF research impact pilot exercise: lessons learned project: Feedback on pilot submission*. London: United Kingdom.

Technopolis (2015). *REF Accountability Review: Costs, benefits and burden*. Technopolis group, Brighton: UK.

Technopolis (2018). *Review of the Research Excellence Framework: Evidence Report*. Technopolis group, Brighton: UK.

Thorpe, A., Craig, R., Hadikin, G., & Batistic, S. (2018). Semantic tone of research 'environment' submissions in the UK's Research Evaluation Framework 2014. *Research Evaluation*, 27(2), 53-62.

Turri, M. (2005) *La valutazione dell'Università. Un'analisi dell'impatto istituzionale e organizzativo*. Milano: Guerini e Associati.

UCAS. (2018). The Teaching Excellence and Student Outcomes Framework (TEF) and demand for full-time undergraduate higher education. Available at: https://www.ucas.com/file/173266/download?token=OVbDbdKZ.

Universities UK (2019). The future of the TEF: report to the independent reviewer. Available at: https://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2019/future-of-the-tef-independent-reviewer.pdf

University and College Union. (2013). *The Research Excellence Framework (REF) - UCU Survey Report.* London: UK.

Van Thiel, S. and Leeuw, F. (2002). The performance paradox in the public sector, *Public Performance and Management Review*, 25(3), 267-281.

Weingart, P. (2005), Impact of Bibliometrics Upon the Science System: Inadvertent Consequences?, *Scientometrics*, 62(1), 117-31.