Type of Evaluation and Features of the Activity to be evaluated: Where does the Trap lie? The Case of the Teaching Excellence Framework

Barbato, Giovanni*

(*) Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Milan, Italy.

Corresponding author: Barbato Giovanni: giovanni.barbato@unimi.it; Department of Economics, Management and Quantitative Methods (DEMM), Università degli Studi di Milano, Via Conservatorio 7, 20122, Milan, Italy.

Type of Evaluation and Features of the Activity to be evaluated: Where does the Trap lie? The Case of the Teaching Excellence Framework

Abstract
The article adopts the Management Control Theory (MCT) perspective to investigate the relationship between the characteristics of tasks/activities to be assessed and the type of assessment employed. Two analytical dimensions are considered: the measurability/attributability of outputs and the knowledge of cause-effect relations producing outputs. The introduction of an evaluation system of teaching in higher education is here used as case-study. The article shows how the complexity of teaching, expressed by a high interdependency among actors and multiple heterogeneous outputs, is not adequately tackled by an evaluation system that is narrowly focused on the quantification of observable outputs. Unintended consequences might therefore arise.

Keywords Performance Evaluation, Performance Indicators, Management Control Theory, Higher Education, Teaching Excellence Framework

1. Introduction

Under the influence of the New Public Management (NPM) movement, performance evaluation (PE) has progressively shaped the operation, management and funding of many public sector organizations. Nevertheless, the impact of PE systems on organizational performance has often been claimed to be unclear, controversial, or even dysfunctional (Adcroft & Willis, 2005; Bevan & Hood, 2006; Diefenbach, 2009; Smith, 1995; Van Thiel & Leeuw, 2002). While several empirical studies have described different types of unintended consequences resulting from the introduction of PE systems, it is also claimed that theoretical reflections on their origin are underdeveloped (Siverbo et al., 2017). In this debate, a dimension that has often been overlooked is the relationship between the type of evaluation instrument employed to assess a specific task/activity and the features of the latter and how these influence one another (Abma & Noordegraaf, 2003; Barbato et al., 2018). NPM indeed provides a specific concept of PE, mainly skewed towards the assessment of measurable and observable outputs via a set of quantitative indicators (Diefenbach, 2009). The apparent appeal of performance indicators (PIs) comes from their ability to measure complex, multifaceted constructs and communicate them

to external stakeholders in an immediate and simple manner while increasing the accountability of those who are evaluated.

However, the literature has shown that many public sector activities are often characterized by low measurability and observability with respect to their outputs (Barbato et al., 2018; Mascarenhas, 1996), a low level of routine and a relevant interdependency between the actors that contribute to carrying out certain tasks/activities (Abma and Noordegraaf, 2003). These peculiarities influence the capacity of an evaluation system to fully capture the performance and quality of public sector activities, shedding light on the importance of the relationship between the nature/features of the task/activity under evaluation and the type of evaluation adopted. An interesting framework that focuses exactly on this connection is provided by management or organizational control theory (MCT) (Eisenhardt, 1985; Frey et al., 2013; Ouchi, 1979). According to MCT, two main analytical dimensions determine which type of evaluation best fits the features of the activity/task to be assessed: the measurability/attributability of the outputs and the knowledge of the cause-effect relations producing the outputs.

The present article aims to contribute to the aforementioned debate by adopting MCT to analyze NPM-driven PE systems introduced in the higher education (HE) sector. The HE sector, similar to other public domains, has indeed been heavily influenced by the NPM narrative, and PE practices have been widely introduced therein (Kallio et al., 2017). Nevertheless, the operation and effects of PE systems in HE have been only superficially compared to those in other sectors (Dal Molin et al., 2017), and the MCT framework has rarely been applied (Minelli et al., 2015; Rebora & Turri, 2013).

The specific PE system considered here is the Teaching Excellence and Student Outcomes Framework (TEF). The TEF is a metrics-based evaluation system first introduced in England in 2016 that operates in tandem with the existing quality assurance (QA) system. The evaluation

of teaching can be a fruitful example for investigating the relationship between the nature of the activity under evaluation and the type of evaluation used because of its complex and ambiguous nature, the presence of multiple outputs with different degrees of measurability and the interdependency of different actors who contribute to the success of teaching activities (Leiber, 2019). Furthermore, while the literature has sufficiently focused on the introduction of PE systems for research (Agyemang & Broadbent, 2015; Rebora & Turri, 2013), studies on the evaluation of teaching have mainly concentrated on QA mechanisms or specific evaluative instruments, such as student surveys (Adcroft & Willis, 2005; Minelli et al., 2015). By comparison, metrics-based PE systems such as the TEF have received less attention.

Therefore, based on the analytical framework provided by MCT, the present article seeks to answer the following research questions:

- How do the specific features/qualities of teaching activities interact with the evaluation system designed to assess them (the TEF)?
- How might potential unintended consequences of the TEF arise?

This article will use secondary sources, such as previous studies and technical reports from both the national regulatory body in charge of implementing the TEF (the Office for Students - OfS) and the Department for Education (DfE) to answer these research questions.

The paper is organized as follows: the next section explains MCT and the analytical dimensions through which PE systems can be analyzed. The third section describes the rationales and operation of the TEF. The fourth section analyzes the TEF through the lens of MCT, focusing on the relationship between the features of teaching as an activity and the characteristics of the TEF as a PE system. Final recommendations and policy implications are provided in the last section.

2. Theoretical Framework

Management control theory (MCT), or organizational control theory, stems from the works of Ouchi (1979) and Thompson (1967), who identified different types of control strategies and showed to which kinds of activities/tasks they apply. MCT indeed claims that to align individuals' efforts with organizations' goals, control systems must be designed according to the nature of the activity/task to be assessed (Eisenhardt, 1985). In this regard, two main analytical dimensions are claimed to be particularly relevant (Ouchi, 1979): 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).

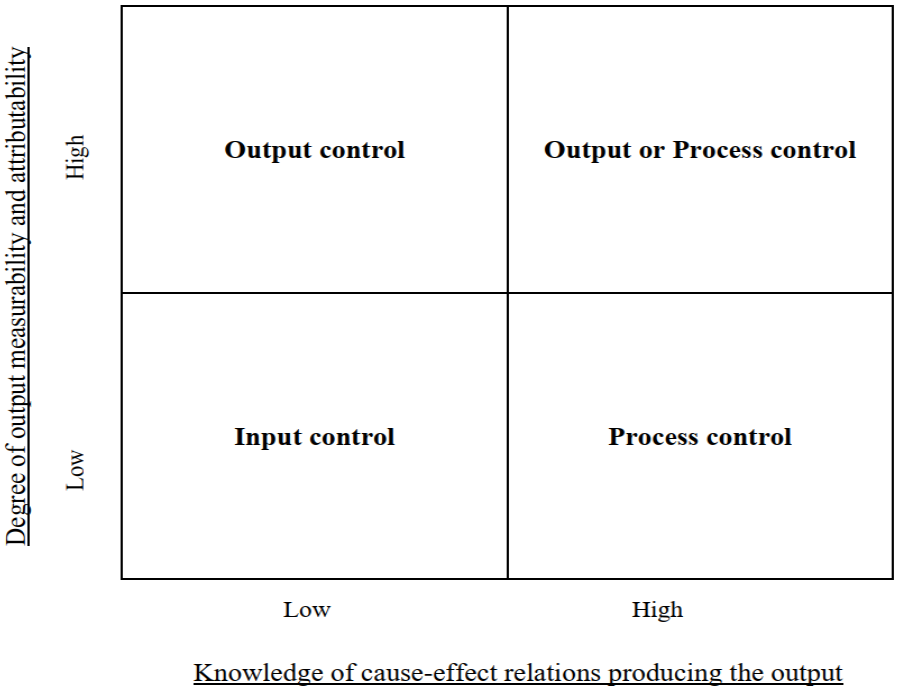
As depicted in Figure 1, the matching of these two analytical dimensions leads to three different types of control. Here, the term control refers to evaluation, as in other works (Frey et al., 2013). The first typology of evaluation is that based on the assessment of outputs. This type of evaluation is effective only when outputs are observable and not 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. Therefore, this kind of evaluation is typical when knowledge of the cause-effect relationship is poorly specified, but output can be easily quantified (Eisenhardt, 1985; Frey et al., 2013).

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 activities/tasks to be evaluated. In this context, evaluation should verify that the 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), 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 for such evaluation are

selection and career procedures and all the processes that socialize and train individuals in the rules and codes of practice 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.

The last typology identified by MCT is process control, which assesses the processes (e.g., standards and guidelines) normally used to achieve certain outputs. This evaluation mechanism can be used when the measurability of outputs is not high but evaluators display sufficient knowledge and shared understanding of the transformation process used to generate them (Eisenhardt, 1985; Turner & Makhija, 2006). Therefore, process control is characterized by an equity of treatment and a transparency in evaluators' judgment that are the result of acceptance of a shared code of practices and behaviors. In this sense, peer review can be considered a form of process control when it is applied for the evaluation of scientific publications (Osterloh, 2010).

Figure 1. Type of evaluation (control) and characteristics of the activities/tasks to be evaluated



Source: Adapted by the author from Ouchi (1979)

Against this background, scholars have widely underlined how NPM-based evaluation systems introduced in the public sector are predominantly skewed towards the assessment of measurable outputs (Frey et al., 2013; Mascarenhas, 1996), a tendency known as "tunnel vision" (Smith, 1995). The primary consequence is that PE does not give proper consideration to more qualitative, value-oriented and ethical aspects that are essential for public goods (Stewart & Walsh, 1994). Most public activities, such as education, health and security, are indeed intrinsically complex, with tasks so highly interdependent that outputs cannot be precisely attributed to individuals (Frey et al., 2013).

Moreover, as underlined by Abma and Noordegraaf (2003), several processes carried out by public administrations are characterized by a low degree of routine and by two-sided interaction, which is when activities are generated by both producers and consumers. Both features naturally reduce the ability of a PE system to truly measure outputs, as also illustrated by other scholars (Frey et al., 2013; Hackman & Oldman, 1980). The tendency to assess only what can be quantified through detailed metrics may ultimately lead to what Meyer and Gupta (1994) describe as a "performance paradox". This phenomenon occurs when the evaluative metrics misrepresent the actual level of performance achieved because they lose their capacity to distinguish between satisfactory and unsatisfactory performance (Barbato et al., 2018; Van Thiel & Leeuw, 2002).

Finally, when PE systems are related to the provision of incentives, unintended consequences can be intensified if the type of evaluation adopted is not aligned with the features of the object/activity to be evaluated. As claimed by Speklé and Verbeeten (2014), if output measurability is low and cause-effect knowledge is ambiguous, incentives will probably induce individuals/organizations to concentrate their efforts only on the activities that are considered in the final evaluation or even on those more easily achievable regardless of their actual

relevance. Similarly, the concept of 'gaming' represents the voluntary manipulation/alteration of the evaluation process, which improves future evaluative judgment without substantially affecting performance. Different gaming practices or other dysfunctional effects have been widely documented across public sectors (Barbato & Turri, 2017; Bevan & Hood, 2006; Diefenbach, 2009; Speklé & Verbeeten, 2014; Van Thiel & Leeuw, 2002) and within the HE sector as well (Adcroft & Willis, 2005; Kallio et al., 2017; Osterloh, 2010).

Therefore, due to its specific focus on the relationship between the type of evaluation and the characteristics of the activities/tasks to be assessed, MCT represents a fruitful theoretical lens through which to analyze NPM-driven PE systems like the TEF.

3. The Teaching Excellence and Student Outcomes Framework (TEF)

The English HE system has a long tradition of evaluating university performance. The first national evaluation exercise of research was introduced in 1986, whereas the Teaching Excellence and Student Outcomes Framework (TEF) was introduced in 2016, but an earlier attempt[1] can be traced back to the 1990s.

The introduction of the TEF was initially mentioned in the Conservative Party's manifesto for the 2015 general election. The manifesto argued that to "ensure that universities deliver the best possible value for money to students: we will introduce a framework to recognize universities offering the highest teaching quality" (Conservative Party, 2015, p. 35). The TEF was thus thought to be an instrument to guarantee that students, who significantly contribute to the funding of the HE sector through high tuition fees, receive an adequate return in terms of the

[1] The Teaching Quality Assessment (TQA) operated from 1993 to 1997 in England and involved a system based on self-assessment, external assessment, and peer review. Three possible assessment outcomes could be given: 'unsatisfactory', 'satisfactory' and 'excellent'. The TQA was ceased due to its lack of transparency and the significant bureaucratic burden it represented for universities.

quality of the services for which they pay. Students are indeed treated as consumers in line with the narrative of the market-based policies that have reformed the English HE system in recent decades (Deem & Baird, 2020; Gunn, 2018). The TEF was thus intended to inform students' decisions about where to study by providing information on the quality of teaching activities at English universities.

By providing incentives for teaching, the TEF was also thought to counterbalance the excessive emphasis of both academics and universities on research brought about by the older research evaluation system (the REF) (Wood & Su, 2017).

The 'market' narrative used to justify the TEF has drawn criticism, since it was claimed to excessively commodify the nature of HE as public good (Deem & Baird, 2020). Moreover, the TEF's quick and top-down implementation by the Department for Education (DfE) and the Minister for Universities, Jo Johnson (Gunn, 2018), also drew criticism. After consultation on its metrics, an introductory year of the TEF was carried out in 2016, and the system was legally implemented through the 2017 Higher Education and Research Act.

3.1 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 quantitative 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[2] (OfS) is the body in charge of coordinating and managing the TEF. In contrast to the REF, the TEF 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 based on two subject-level pilot exercises conducted in a sample of universities. The results of these pilot exercises will also inform

[2] As a result of the 2017 Higher Education and Research Act, the Office for Students has replaced HEFCE as the regulatory body for teaching in higher education.

Shirley Pearce's independent review of the TEF, which is expected to provide recommendations in 2020, leading to a potentially 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 also involved.

The evaluation process is based on two main sources of evidence: the quantitative indicators reported in Table 1 and a qualitative and narrative document known as the 'provider submission'.

Table 1. TEF core and supplementary metrics, description and the data source

Areas	**Core and *supplementary* metrics**	**Description and source**
Teaching quality (1)	Student satisfaction with teaching on their courses (1a)	National Student Survey (NSS), questions 1 to 4 (at 2018): The metric is built on the basis of four questions regarding how good teaching staff are at explaining and making the subject interesting and how the course has challenged students to achieve their best work.
	Student satisfaction with assessment and feedback (1b)	National Student Survey (NSS), questions 8 to 11 (at 2018): The metric is developed on the basis of four questions about the clarity and fairness of marking criteria as well as the utility and promptness of feedback from teaching staff to students.
	Grade inflation (1c)	HESA[3] and ILR[4] student records: This metric provides information on the types of degrees awarded (1sts, 2:1s, other degree classifications and unclassified degree awards)
Learning Environment (2)	Student satisfaction with academic support (2a)	National Student Survey (NSS), questions 12 to 14 (at 2018): This metric is built on the basis of three questions on the possibility of asking teaching staff for advice/guidance and the quality of such advice.
	Student retention on courses (2b)	HESA and ILR student records: This indicator tracks students from the year they enter a HE provider to the next academic year. It verifies whether students are recorded as actively studying in a HE program.
Student outcomes (3)	Employment or further study (3a)	DLHE[5] survey: Percentage of UK-domiciled leavers who are working or continuing their study 6 months after graduation.
	Highly skilled employment or further study (3b)	DLHE survey: Percentage of UK-domiciled leavers who are in highly skilled employment[6] or studying at 6 months after graduation.
	Above median earnings (3c)	LEO[7] dataset:

[3] Higher Education Statistical Agency (HESA).
[4] Individualized Learner Record (ILR).
[5] Destinations of Leavers in Higher Education (DLHE).
[6] The UK Standard Occupational Classification (SOC) classifies jobs within groups 1, 2 and 3 as 'highly skilled'.
[7] Longitudinal Education Outcomes (LEO) 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 continuing their study three years after graduation.

Source: Adapted from Office for Students (2018) and Department of Education (2017)

The quantitative indicators are divided into six core metrics and three supplementary metrics. The metrics related to 'teaching quality' (1a and 1b) and one related to 'learning environment' (2a) are measures of students' satisfaction calculated based on a set of questions from the National Student Survey[8] (NSS). The other indicators regard employment status (3a) and type of graduates' employment (3b) as well as the regularity of students' careers (2b). As can be noted from Table 1, almost all the metrics are measures of outputs or outcomes (career progression; employability) of learning and teaching activities, while processes and inputs are less considered and only through the perspectives of students (e.g. feedback 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 on the basis of certain 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 based on national HE databases 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 university, such as the entry qualifications of

[8] The NSS records students' opinions on several aspects of their degree programs in the final year of their academic career.

students and the subject of study. Both the benchmark and the difference thus inform assessors how to interpret the metric values and are unique for each university.

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 performance and illustrate their institutional approach towards teaching excellence and how this affects students (OfS, 2018). In this regard, based on an analysis of a sample of HE provider submissions, Beech (2017) showed that additional qualitative and quantitative data are often reported, such as citations from external QA reviews, student union statements, 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 look only at the core metrics, with attention to their distance from the benchmarks (the flags), and use split metrics and contextual data when necessary. The three metrics based on NSS data have a weight of 0.5, while the others equal 1. Based on this process, an initial hypothesis on the rating is generated[9]. Second, the provider submission and the supplementary metrics are then considered to decide if the initial hypothesis can be confirmed or needs to be modified (the second step). Both the first and the second assessment steps are carried out within small groups of panel members that consist of (at least) two academics and (at least) one student. Each group looks at a set of universities. Finally (third step), a meeting of the full TEF panel collectively determines the final rating (a 'Gold', 'Silver' or 'Bronze' medal) based on the recommendations advanced by each group of panel members. A statement of the findings is also provided to each university, in which the reasons behind the rating are explained. Although HE provider submissions can potentially play an important role during the three-step

[9] So, for example, if a university presents a total value of 2.5 based on its core metrics, it should be awarded, at the end of the first step, with the 'Gold' rating.

assessment process, it has been highlighted that only 15% of initial hypotheses are changed after analysis of the provider submission (Matthews & Kotzee, 2020).

4. The TEF under the Analytical Lens of the Management Control Theory

As illustrated in section 2, MCT provides an analytical framework through which the relationship between the features of the activity/task to be assessed and the most appropriate type of evaluation for this assessment can be examined. This section employs the two analytical dimensions provided by MCT, namely, the degree of the measurability/attributability of outputs and the knowledge of the cause-effect relations producing the outputs, to analyze the most relevant features of teaching and how the TEF manages (or does not manage) to take them into account.

First, teaching activities do not have just one single and measurable output, as may appear to be the case with scientific publications for research, since many outputs might be recognized as such. Some outputs can be evaluated through single performance indicators (PIs), such as the employability of graduates or student dropout, while others are not as easy to capture through punctual metrics (Leiber, 2019; Tam, 2001). An example of a fuzzier and more qualitative output is certainly student learning gain, which requires sophisticated evaluative mechanisms, such as the comparison of knowledge/skills before and after different learning phases, to ultimately be quantified.

Second, teaching activities are characterized by an intensive interdependency between teachers and students, which ultimately affects the success of these activities (Wood & Su, 2017). Teaching can thus be claimed to be characterized by a two-sided interaction as described by Abma and Noordegraaf (2003). The success of teaching is certainly shaped by teaching competences/skills displayed by the teacher, but it also depends on students' attitudes and

efforts. These last are mainly the result of prior educational paths and personal and intellectual capacities that are not necessarily determined by greater teacher effort. This interdependency results in the unclear attributability of outputs: are teaching outputs the effect of teachers' efforts alone? How important are students' attitudes and backgrounds in making teaching processes effective? These are questions that cannot be taken for granted and that suggest, in terms of assessment, that both the teacher and student sides should be jointly considered during evaluative processes (Tam, 2001; Wood & Su, 2017).

Moreover, the mutual and unpredictable connection between teachers and students makes the knowledge of the cause-effect relations imperfect and partial for the evaluator. This situation undermines the preconditions for using process-based types of evaluation (Frey et al., 2013), since only strict observation of teaching and learning processes can improve the know-how needed to assess them comprehensively. However, this would also entail potential drawbacks represented by additional bureaucratic and financial costs, as is often claimed in relation to peer review-based systems like the REF (Geuna & Piolatto, 2016).

However, as underlined by Gibbs (2008), teaching quality does not stem solely from student-teacher interaction, since environmental factors such as curriculum design, services to students and the quality of infrastructure matter.

In summary, this partial measurability and attributability of outputs and nonlinear knowledge of the cause-effect relations leading to the outputs suggests that the evaluation of a few quantitative outputs might not be sufficient to capture the complex nature of teaching.

Nevertheless, as illustrated in the third section, the TEF is strongly oriented towards the assessment of a narrow set of teaching outputs and outcomes (Gunn, 2018; Wood & Su, 2017;), covering only those that are easily measured through quantitative measures, namely, student satisfaction and graduates' employability, resulting in the abovementioned issue of "tunnel vision" (Smith, 1995). This tendency has also been registered in other HE contexts (Liu, 2015)

and means that other more qualitative but still crucial aspects of teaching, such as learning gains, teachers' competences/attitudes, learning strategies and curriculum design (Cui et al., 2019; Leiber, 2019), are omitted from the assessment process. Furthermore, it is also relevant to emphasize that while students' feedback on teaching quality is given considerable weight, teachers' viewpoint is basically absent from the TEF metrics, resulting in a failure to consider both sides of the coin. Consequently, as underlined by Siverbo et al. (2019), the PE system is incomplete and thus provides a partial representation of performance.

Additionally, the more qualitative HE provider submissions do not seem to particularly enrich the knowledge of teaching performance already depicted by the quantitative metrics. Matthews and Kotzee (2020) empirically investigated, through text analysis, the HE provider submissions of those universities that were upgraded after the first step (based on the interpretation of the TEF metrics). The authors found that the themes in the HE provider submission texts that received the most attention were employment, employability and learning outcomes, which clearly overlap with both the quality criteria and the coverage of the quantitative metrics. Therefore, "Successful submissions followed, in some ways quite literally, a 'script' and self-consciously mirrored the language of the TEF as a bureaucratic exercise" (Matthews & Kotzee, 2020, p. 18) and did not take the opportunity to highlight other crucial aspects of teaching activities that have been neglected by the quantitative metrics.

Second, TEF metrics (Table 1) are anything but uncritical. Concerning the student satisfaction metrics (1a, 1b, 2a), it is claimed that students' satisfaction and teaching quality are different constructs, since satisfaction is influenced by factors outside of the teaching process itself (Spooren et al., 2013). No claim is being made here that students' satisfaction is unrelated to teaching quality. However, it can be argued that metrics built on student surveys can serve as markers but are not able to comprehensively represent a multilayered concept such as teaching (Wood & Su, 2017). Metrics regarding the employment of graduates (3a and 3b) are also

claimed to be partially affected by factors that do not depend on universities' efforts in teaching, as they are related, e.g., to the health of the local labor market and economy and the discipline itself (Deem & Baird, 2020; UUK, 2019). The ambiguity regarding which factors ultimately affect the teaching outputs considered here (employability and student satisfaction) as well as their incomplete relationships with teaching activities might therefore weaken the ability of the TEF metrics to discriminate between good and poor performance.

Third, the 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 argued that the TEF measures teaching quality only indirectly, with questionable metrics and with a narrow perspective (O'Leary & Wood, 2019; UUK, 2019), resulting in a partial representation of teaching quality within universities.

As suggested by MCT scholars, this imbalance between the features of the object of assessment and the type of evaluation adopted may lead to the emergence of unintended consequences. Although evidence on the effects of the TEF on academics and universities is still limited, some potential risks might already be envisaged.

First, since the TEF metrics do not fully capture all teaching dimensions, university management might implement strategies and invest resources with the narrow goal of improving the activities measured by these metrics (e.g., employability), thus losing a more holistic vision of learning and teaching processes. This unintended consequence is known as "measure fixation" (Smith, 1995, p. 290). The first empirical evidence on the TEF seems to point to such tendencies. Cui et al. (2019) illustrate how the TEF has certainly increased the internal centralization and standardization of teaching activities as well as the accountability of academics, with activities directed mainly at satisfying TEF metrics and not at improving the overall L&T experience.

Second, the high relevance of student satisfaction metrics could provide negative incentives for universities to discourage innovative forms of teaching, since "they often score low student satisfaction ratings, despite these methods often being highly effective in enhancing student learning" (RSS, 2016, p. 1). Similar arguments are supported by the empirical inquiry of Sutherland et al. (2018) and by Kallio et al. (2017, p. 299) in a study on the Finnish reality: "The easiest way of meeting targets is by lowering quality, for instance by letting students pass exams more easily and granting degrees with looser criteria". Therefore, a partial representation of performance might lead to behavioral displacement among those who are influenced by the evaluation, leading to an opposite result than expected (Siverbo et al., 2017).

5. Concluding remarks

The present paper adopted the analytical framework of MCT and the case of the TEF to illustrate how the relationship between the specific features of an activity/task to be evaluated and the type of evaluation employed can be overlooked and cause potential unintended consequences, confirming previous studies (Barbato & Turri, 2017; Frey et al., 2013; Speklé & Verbeeten, 2014). This relation seems particularly relevant for those public activities, such as teaching, that are characterized by outputs with low measurability/attributability and imperfect knowledge of the processes leading to these outputs due primarily to the multitude of outputs and the high interdependency between the main subjects involved (teachers and students). When these features are neglected, it follows that PE systems provide a partial representation of the activity to be evaluated (Frey et al., 2013), as emerged from the analysis of the TEF. Regarding the specific case of teaching evaluation, three main lessons can be formulated for both policymakers and scholars.

First, to effectively evaluate teaching, a more holistic approach that is able to represent all the relevant dimensions of this complex activity must be adopted. In this regard, Leiber (2019, p.

79), in presenting a comprehensive list of 230 PIs from two research projects (QUELIT[10] and SQELT[11]), claims that four subdomains of L&T should be jointly considered during evaluation procedures: "L&T environment, Teaching processes and competences of teachers, Learning processes and competences of learners; Learning outcomes and learning gain and their assessment". PIs could then be developed for each subdomain to represent the inputs, processes and outputs/outcomes of L&T. A similar approach has also been presented in Chalmers (2008). Notably, the literature has highlighted that the real challenge in the evaluation of teaching is the shift from the assessment of teaching to that of the learning experience and gains, thus putting students at the center (Barr & Tagg, 1995; Wood & Su, 2017). An example of a recent attempt made in this direction is certainly the AHELO project, which focuses on the measurement of the skills and competences of students who have completed their bachelor's degree (Dias & Amaral, 2014).

A more comprehensive approach towards the evaluation of teaching should also be expressed in how the assessment results are communicated to the main stakeholders. Regarding this point, the ranking and medal-based system of the TEF seems particularly arid even though it was also introduced to inform students. A survey carried out by the Universities and Colleges Admissions Service (UCAS) (2018) indeed highlights that only 17.1% of students approaching HE know what the TEF is, and half of them state that is was useful in deciding which university to choose.

Second, data on teaching activities are rarely collected in a systematic and comparable way (Sarrico et al., 2010). This is partly due to the intrinsic aforementioned difficulty in capturing all of the outputs of L&T processes but also because universities and academics' performance has been evaluated predominantly in terms of research quality without proper attention to teaching (Gunn, 2018). It is thus necessary that both scholars and policy-makers deepen their

[10] On the Way to Sustainable Quality Enhancement in Learning and Teaching
[11] Sustainable Quality Enhancement in Higher Education Learning and Teaching

knowledge on how to measure more qualitative and procedural aspects of teaching activities (Leiber, 2019). In this regard, some studies have underlined that the TEF has contributed to rebalancing attention in favor of teaching, even though the majority of academics (and universities) still say that the effect on the actual improvement of teaching quality has been very small (Cui et al., 2019; UUK, 2019).

Finally, the MCT underlines that when the measurability of outputs and knowledge of the cause-effect process are particularly low, an evaluation of the inputs can partly reduce the distortive effects of a pure output-based assessment (Turner & Makhija, 2006). Regarding teaching activities, this insight might be interpreted as a call to increase attention to the potential value of faculty development and pedagogical training for academics in the early career stage. This point might be particularly relevant in HE systems where career advancement is based only on research performance and young academics are trained only to become researchers.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Abma, T. A., & Noordegraaf, M. (2003). Public managers amidst ambiguity: Towards a typology of evaluative practices in public management. *Evaluation*, 9(3), 285–306. https://doi.org/10.1177/13563890030093004.

Adcroft, A., & Willis, R. (2005). The (un)intended outcome of public sector performance measurement. *International Journal of Public Sector Management*, 18(5), 386-400. https://doi.org/10.1108/09513550510608859.

Agyemang, G., & Broadbent, J. (2015). Management control systems and research management in universities: An empirical and conceptual exploration. *Accounting, Auditing & Accountability Journal*, 28(7), 1018-1046. https://doi.org/10.1108/AAAJ-11-2013-1531.

Barbato, G., & Turri, M. (2017). Understanding public performance measurement through theoretical pluralism. *International Journal of Public Sector Management*, 30(1), 15–30. https://doi.org/10.1108/IJPSM-11-2015-0202.

Barbato, G., Salvadori, A., & Turri, M. (2018). There's a lid for every pot! The relationship between performance measurement and administrative activities in Italian ministries. *Cogent Business & Management*, *5*(1), 1-20. https://doi.org/10.1080/23311975.2018.1527965.

Barr, R. B., & Tagg, J. (1995). From teaching to learning. A new paradigm for undergraduate education. *Change: The Magazine of Higher Learning*, 27(6), 12–26. https://doi.org/10.1080/00091383.1995.10544672.

Beech, D. (2017). *Going for gold: Lessons from the TEF provider submissions* (Report No. 99). Higher Education Policy Institute (HEPI)

Bevan, G., & 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. https://doi.org/10.1111/j.1467-9299.2006.00600.x.

Chalmers, D. (2008, September 8-10). *Teaching and learning quality indicators in Australian universities* [Paper presentation] Institutional Management in Higher Education Conference, Paris, France.

Cui, V., French, A., & 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*. Advance online publication. https://doi.org/10.1080/03075079.2019.1704721.

Dal Molin, M., Turri, M., & Agasisti, T. (2017). New public management reforms in the Italian universities: Managerial tools, accountability mechanisms or simply compliance?. *International Journal of Public Administration*, 40(3), 256-269. https://doi.org/10.1080/01900692.2015.1107737.

Deem, R., & Baird, J. (2020). The English teaching excellence (and student outcomes) framework: Intelligent accountability in higher education?. *Journal of Educational Change*, 21, 215-243. https://doi.org/10.1007/s10833-019-09356-0.

Department for Education (DfE) (2017). *Teaching excellence and student outcomes framework specification*.

Dias D., & Amaral A. (2014). Assessment of higher education learning outcomes (AHELO): An OECD feasibility study. In M. J. Rosa & A. Amaral (Eds.), *Quality assurance in higher education. Contemporary debates* (pp. 66-87). Palgrave Macmillan.

Diefenbach, T. (2009). New public management in public sector organizations: The dark sides of managerialistic enlightenment. *Public Administration*, 87(4), 892-909. https://doi.org/10.1111/j.1467-9299.2009.01766.x.

Eisenhardt, K. M. (1985). Control: Organizational and economic approaches. *Management Science*, 31(2) 134–149. https://doi.org/10.1287/mnsc.31.2.134.

Frey, B. S., Homberg, F., & Osterloh, M. (2013). Organizational control systems and pay-for-performance in the public sector. *Organizational Studies*, 34(7), 949-972. https://doi.org/10.1177/0170840613483655.

Geuna, A., & Piolatto, M. (2016). Research assessment in the UK and Italy: Costly and difficult, but probably worth it (at least for a while). *Research Policy*, 45(1), 260-271. https://doi.org/10.1016/j.respol.2015.09.004.

Gibbs, G. (2008). *Conceptions of teaching excellence underlying teaching award schemes*. Higher Education Academy.

Gunn, A. S. (2018). Metrics and methodologies for measuring teaching quality in higher education: Developing the teaching excellence framework (TEF). *Educational Review*, 70(2), 129–148. https://doi.org/10.1080/00131911.2017.1410106.

Hackman, J. R., & Oldham, G. R. (1980). *Work redesign*. Addison-Wesley.

Kallio, K. M., Kallio, T. J., & Grossi, G. (2017). Performance measurement in universities: Ambiguities in the use of quality versus quantity in performance indicators. *Public Money & Management*, 37(4), 293-300. https://doi.org/10.1080/09540962.2017.1295735.

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. https://doi.org/10.1080/13538322.2019.1594030.

Liu, S., (2015). Higher education quality assessment in China: An impact study. *Higher Education Policy*, 28(2), 175-195. https://doi.org/10.1057/hep.2014.3.

Matthews, A., & Kotzee, B. (2020). The rhetoric of the UK higher education Teaching Excellence Framework: a corpus-assisted discourse analysis of TEF2 provider statements. *Educational Review*. Advance online publication. https://doi.org/10.1080/00131911.2019.1666796.

Mascarenhas, R. C. (1993). Building an enterprise culture in the public sector: Reform of the public sector in Australia, Britain, and New Zealand. *Public Administration Review*, 53(4), 319-328. https://doi.org/10.2307/977144.

Meyer, M. W., & Gupta, V. (1994). The performance paradox. *Research in Organizational Behavior*, 16, 309–369.

Minelli, E., Rebora, G., & Turri, M. (2015). Quest for accountability: Exploring the evaluation process of universities. *Quality in Higher Education*, 21(2), 103-131. https://doi.org/10.1080/13538322.2015.1066611.

Office for Students (OfS) (2018). *Teaching excellence and student outcomes framework (TEF). Year four procedural guidance* (OfS 45/2018).

O'Leary, M., & 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. https://doi.org/10.1080/00131911.2019.1524203.

Osterloh, M. (2010). Governance by numbers: Does it really work in research?. *Analyse und Kritik*, 32(2), 267–283. https://doi.org/10.1515/auk-2010-0205.

Ouchi, W. (1979). A conceptual framework for design of organizational control mechanism. *Management Science*, 25(9), 833-848. http://www.jstor.org/stable/2630236.

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

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

Sarrico, C. S., Rosa, M. J., Teixeira, P. N., & Cardoso, M. F. (2010). Assessing quality and evaluating performance in higher education: Worlds apart or complementary views?. *Minerva*, 48(1), 35-54. https://doi.org/10.1007/s11024-010-9142-2.

Siverbo, S., Cäker, M., & Åkesson, J. (2019). Conceptualizing dysfunctional consequences of performance measurement in the public sector. *Public Management Review*, 21(12), 1801-1823. https://doi.org/10.1080/14719037.2019.1577906.

Smith, P. (1995). On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration* 18(2-3), 277-310. https://doi.org/10.1080/01900699508525011.

Speklé, R. F., & 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. https://doi.org/10.1016/j.mar.2013.07.004.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching. *Review of Educational Research*, 83 (4), 598–642. https://doi.org/10.3102/0034654313496870.

Sutherland, D., Warwick, P., Anderson, J., & Learmonth, M. (2018). How do quality of teaching, assessment, and feedback drive undergraduate course satisfaction in UK business schools? A comparative analysis with nonbusiness school courses using the UK national student survey. *Journal of Management Education*, 42(5), 618-649. https://doi.org/10.1177/1052562918787849.

Tam, M. (2001). Measuring quality and performance in higher education. *Quality in Higher Education*, 7(1), 47-54. https://doi.org/10.1080/13538320120045076.

Thompson, J. D. (1967). *Organizations in action: Social science bases of administrative theory*. McGraw-Hill Book Company.

Turner, K. L., & Makhija, M. V. (2006). The role of organizational controls in managing knowledge. *Academy of management review*, 31(1), 197-217. https://doi.org/10.2307/20159192.

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

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

Van Thiel, S., & Leeuw, F. (2002). The performance paradox in the public sector. *Public Performance and Management Review*, 25(3), 267-281. https://doi.org/10.1080/15309576.2002.11643661.

Wood, M., & Su, F. (2017). What makes an excellent lecturer? Academics' perspectives on the discourse of 'teaching excellence' in higher education. *Teaching in higher education*, 22(4), 451-466. https://doi.org/10.1080/13562517.2017.1301911.