Omega 40 (2012) 424–436

Contents lists available at SciVerse ScienceDirect

Omega

journal homepage: www.elsevier.com/locate/omega

A multicriteria decision analysis model for faculty evaluation

Carlos A. Bana e Costa, Mónica D. Oliveira *

CEG-IST, Centre for Management Studies of Instituto Superior Técnico, Technical University of Lisbon, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

ARTICLE INFO

Article history:
Received 16 November 2010
Accepted 25 August 2011
Processed by Yeh
Available online 7 September 2011

Keywords:
Multicriteria decision analysis
Faculty evaluation
Decision support systems
Higher education

ABSTRACT

In the context of increasing demands for social and financial accountability of universities, the required implementation of transparent faculty evaluation systems constitutes a challenge and an opportunity for universities strategically aligning the activity of academic staff with the university goals. However, despite growing interest in the performance appraisal of faculty, only a few reported studies propose models that cover the full range of academic activities and the models in use are typically based on *ad hoc* scoring systems that lack theoretical soundness. This article approaches faculty evaluation from an innovative comprehensive perspective. Based on the concepts and methods of multiple criteria value measurement, it proposes a new faculty evaluation model that addresses the whole range of academic activities and can be applied within and across distinct scientific areas, while respecting their specificities. Constructed through a socio-technical process, the model was designed for and adopted by the Instituto Superior Técnico, the engineering school of the Technical University of Lisbon. The model has a two-level hierarchical additive structure, with top-level evaluation areas specified by second-level evaluation criteria. A bottom non-additive third level accounts for the quantitative and qualitative dimensions of academic activity related to each evaluation criterion. The model allows (a) the comparison of the performance of academic staff with performance targets reflecting the strategic policy concerns of university management; (b) the definition of the multicriteria value profile of each faculty member at the top level of the evaluation areas; (c) the computation of an overall value score for each faculty member, through an optimisation procedure that makes use of a flexible system of weights and (d) the assignment of faculty members to rating categories.

1. Introduction

In the past years there have been considerable changes in the university system of organisation and funding. The traditional activities of teaching, research and service are increasingly committed to the needs of society [1] and universities have been assuming active responsibilities within the economy [2]. In addition, the institutional and legal setting in which many universities operate has undergone major transformations and a global trend towards increasing social and financial accountability of universities is being observed [3]. Bringing faculty evaluation in line with the changes in the university system has become a priority in many countries around the world. Faculty evaluation is becoming more formal and complex, and several associations in the USA have recommended clarity in standards and procedures, consistency over time among candidates with similar profiles, candour in the evaluation of tenure-track faculty and care for unsuccessful candidates [3]. In Europe, the need for developing evaluation tools is recognised both at the national level and at the EU supra-national level [4]. For example, in Spain, national rules have been defined in recent years for the evaluation of academic staff [5]. In Portugal, the universities are presently defining faculty evaluation processes [6].

As a result of these developments, there is a challenge and an opportunity for each university to align the activity of its faculty members with its mission and strategic plans. Universities are expected to make decisions on recruiting, promoting, granting tenure and rewarding excellence based on putative objective evaluation criteria and supported by appropriate tools. However, despite the international growing interest in the performance appraisal of university activities, and in particular in faculty evaluation, there are only a few studies that attempt to evaluate the overall activity of the academic staff [7] and the "existing metrics do not capture the full range of activities that support and transmit scientific ideas" [8] (p. 488). Hence, there is a need to develop comprehensive evaluation systems, based on methodologically sound procedures that can adequately reflect the differences between the academic staff, taking into account the

* Correspondence to: Instituto Superior Técnico, Departamento de Engenharia e Gestão, Avenida Rovisco Pais, 1049-001 Lisboa, Portugal. Tel.: +351 218417322; fax: +351 218 417 979.
E-mail addresses: carlosbana@ist.utl.pt (C.A. Bana e Costa), monica.oliveira@ist.utl.pt (M.D. Oliveira).

0305-0483/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.omega.2011.08.006

university mission, and that are applicable to all faculty members and scientific areas while respecting their specificities.

This paper proposes an innovative model for faculty evaluation, based on concepts and methods of multiple criteria value measurement with strong theoretical foundations (see for example [9,10]). The proposed model is capable of addressing the multidimensional nature of the evaluation problem – where different evaluation components need to be taken into consideration – and flexible enough to integrate both quantitative and qualitative dimensions, in line with recommendations and guidelines on how to build comprehensive faculty evaluation models [11,12]. The model was designed within the legal and institutional context of the Portuguese universities to be used by the Instituto Superior Técnico (IST) of the Technical University of Lisbon (TUL). IST is an engineering school with 778 faculty members working in a wide variety of scientific domains (ranging from mathematics, physics and chemistry to most branches of engineering, architecture and management).

Section 2 presents briefly the state of the art in the faculty evaluation literature, Section 3 presents the features of the adopted multicriteria modelling approach, Section 4 describes how the multicriteria approach was developed at IST and, finally, Section 5 discusses what was achieved and what is still ahead.

2. Background on faculty evaluation

Personnel management, self-improvement, the growth and development of faculty members and the improvement of the quality of instruction in schools are understood to be the key objectives for faculty evaluation [13]. Given the nature of academic activity and the organisational structure of universities, evaluation systems of academics in use in universities are mostly based on peer reviews. Nevertheless, differences exist in the information basis and methods that peers might use in the evaluation process. While several authors sustain that it is possible to measure faculty performance with some precision and that performance measurements might be used in university management [11], others consider that scientific activities cannot be fully measured given the current knowledge and the available indicators, and that the use of measurement tools might affect researchers' autonomy and might lead to undesirable effects [14]. The different opinions are partly explained by methodological difficulties related to the following:

- It is hard to measure an individual faculty member's total contribution to the school, and the proper balance among research, teaching and service has not been definitely established for the personnel of any type of university [15]. Differing values given to these activities are apparently neither appreciated nor systematically communicated [15]. It is difficult to define which activities to include in scholarship [3] and to find appropriate indicators for performance measurement [14]. Evaluation methods are sensitive to the selected indicators and to the data sources [16].
- Faculty evaluation models typically make use of objective approaches and/or subjective approaches [17]. Objective approaches do not depend on the evaluator (for example, using citation counts) and might generate unintended results because of problems with the data (such as with bibliometric data) [8], generating biases in the evaluation [2]. Subjective approaches, on the other hand, can be influenced by personal biases or by some lack of or insufficient knowledge or experience by some group members [17]. There has been little research on how to integrate objective and subjective approaches adequately [17].
- Given that faculty evaluation implicitly incorporates many beliefs about academic careers and institutional policy, generates different costs and shapes the power relationships between stakeholders, as well as interacts with the balance between personal and departmental goals in academia [15], it is not an easy task to build and promote changes in evaluation systems [18].
- The faculty evaluation literature is spread across several areas. While some professions have held extensive discussions about evaluation models and tools (this being the case of the pharmacy and accounting communities [15,19]), there has been undervaluing or underreporting of research for some communities (e.g. social sciences) [20]. Most evaluation studies explicitly state their area of applicability.

Although to date no movement has emerged to standardise the evaluation process and maximise objectivity while linking productivity in an empirical fashion to rewards [7], multiple institutions have advocated the need to develop an evaluation culture in university systems [4] and to create more comprehensive evaluation systems. This is the case of the National Academy of Engineering in the US [12] and the director of the Science of Science & Innovation Policy programme from the National Science Foundation in the US [8].

An analysis of the evaluation literature in the university context shows that most studies reported carried out comparative analyses of universities, faculties, departments or research units (such as [2,21,22,23]), while only a few propose methods to evaluate academic staff. Nevertheless, it is recognised that faculty members are the ground unit of the academic system, the key unit for analysing university production and an operational unit for the management of human resources (for instance, with respect to promotions).

Most studies on faculty evaluation use qualitative methods to structure the evaluation problem [13,18,19]. Some propose conceptual frameworks and multiple approaches for faculty evaluation [11,21]. To our knowledge, very few studies have used decision analysis models to analyse thoroughly the academic research outputs of individuals [24,25]. However, as far as we are aware, the literature in the area does not provide comprehensive models for the evaluation of academic staff. The literature available on validation methods for students to assess the performance of their teachers, which may lead to payment awards in universities in some countries, including the US [11], is only able to capture a small part of the daily activities of the academic staff and definitely does not cover their performance in research, services and management.

There are many evaluation studies of university units and programs; however "most of the evaluation methodologies used in these studies suffer major flaws in both substance and process" [26]. This also applies to methods used in faculty evaluation like point systems [5], which may incur in well-known mistakes reported in the decision analysis literature, including treating performance indicators as evaluation criteria, not distinguishing between the notion of performance and the notion of value; weighting criteria solely on the basis of the intuitive notion of importance [27], ignoring the notion of value trade-offs underlying additive aggregation models (Keeney [28] calls this the most common critical mistake); and summing up ordinal scores on the criteria giving rise to meaningless overall scores. Also, as remarked by Billaut et al. [29] when reviewing methods used to rank universities, the "… main conclusions are that the criteria that are used are not relevant, that the aggregation methodology is plagued by a number of major problems and that the whole exercise suffers from an insufficient attention paid to fundamental structuring issues" (p. 1).

3. Multicriteria modelling of faculty evaluation

3.1. Methodological framework

A novel faculty evaluation model was designed for IST by an internal working group (WG) of professors, in a sequence of decision conferencing workshops [30,31] in which we acted as impartial facilitators and decision analysts following a process consultation approach [32]. This is a socio-technical interactive and learning model-building process [33,34] integrating the technical elements of multicriteria value measurement [10,33] and the social elements of decision conferencing. The WG was led by the chairman of the Scientific Board of the school assisted by several other Board members that developed a shared understanding of the key faculty evaluation aspects. This was crucial to build the model that the WG proposed to the school decision-maker (DM), which is the Management Board of IST. Subsequently, the DM approved the model in general terms, submitted it to the faculty for discussion and made the final decision to use the model after some adjustments to take into account relevant suggestions. The final step of the implementation process was the designation of a full professor as the evaluator for each faculty member in two evaluation periods: 2004–2007 and 2008–2009.

The model-building process aimed to provide answers to key questions not previously addressed in a comprehensive and systematic manner within the faculty evaluation literature, namely

- When structuring the faculty evaluation model, i.e.
  - How to design a model reflecting the strategic objectives of the school and useful for human resources management?
  - How to define a coherent set of evaluation criteria projecting, in the various areas of academic activity (pedagogical, scientific, etc.), stakeholders' values and concerns about academic careers and institutional policies?
  - How to describe, as objectively and unambiguously as possible, the performance on each one of the criteria, taking into account and adequately integrating its quantitative and qualitative dimensions?
  - How to care for the specificities of each of the scientific domains of the school?
- When modelling the measurement of academics' value within each evaluation criterion and across criteria within each area of activity:
  - How to convert individual performance into perceived added value to the school?
  - How to assign relative weights to the criteria, adequately reflecting value trade-off judgements between criteria?
  - How to aggregate added value on multiple criteria appropriately, within and across areas of activity, respecting the autonomy of each faculty member to choose to invest more in some activities than in others, while not allowing extreme performance compensation phenomena inconsistent with achieving an adequate balance among objectives?
  - How to set boundaries for the rating categories imposed by law, so that the classification of each faculty member may reflect her or his intrinsic value to the school?

The model-building process can be described as a package of entangled activities, developed by the WG during the decision conferences, as detailed in Fig. 1 and discussed in the next subsections.

From the discussion of the methodological and contextual issues related to faculty evaluation, the boundaries of the problem were established and a multicriteria modelling approach could be adopted. Firstly, in structuring the model, criteria were specified to evaluate faculty members within each of the areas of academic activity that are normatively defined in the legislation [6]. Secondly, descriptors of quantitative and qualitative performance

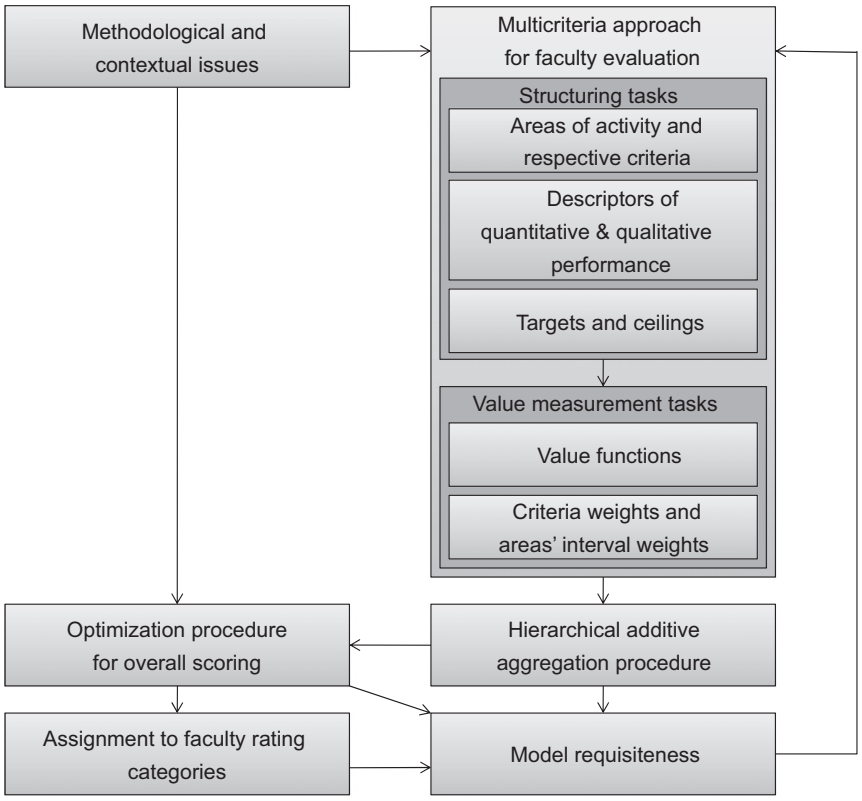

Fig. 1. Key components of the process of building a multicriteria model for faculty evaluation.

for each criterion were defined. Then, the concept of a "target" associated with each criterion was introduced to make it possible to incorporate into the model reference levels of performance reflecting policy concerns. Targets make explicit what should be considered a realistic "good performance" on each criterion for a given scientific domain and in a given evaluation period. This makes it clear that performance is one thing and its value is another.

Measures of value were subsequently built, that is, value functions enabling performance to be transformed into value at the level of each criterion separately. Finally, the areas of activity and the respective criteria were weighted. These weights should reflect the relative importance of achieving the targets, in a given evaluation period, from the perspective of the school.

An additive value procedure could then be applied hierarchically, firstly to aggregate value scores on criteria within each area and then to aggregate values across areas. The area value scores obtained by a faculty member form her or his multicriteria value profile at the top level of the areas of activity. Once different types of "good" academic profiles are not only admitted but also desirable, from a functional perspective, in a faculty body, all types contributing to the achievement of the university strategic objectives, the area weights should be allowed to vary within reasonable bounds. Therefore, interval weights were defined for the areas and an optimisation procedure was adopted [16]. This enables the application of the additive model at the top area level, in such a way that the overall score resulting for each faculty member reflects the value of her or his specific profile. To prevent a very high performance level in a single criterion playing an excessive undesirable role in compensating for very weak performance in all the remaining criteria, the concept of a "ceiling" was introduced into the model. Finally, an assignment procedure makes it possible to associate each faculty member with one rating category, with the several categories separated by boundaries of increasing overall value (which can be combined or not with other assignment rules).

The model will be considered "requisite" (see [34,35]) when its form and content are sufficient to provide satisfactorily uncontroversial answers to the questions that motivated its development. This explains the recursive nature of the scheme in Fig. 1.

There are several theoretically sound methods proposed in the decision analysis literature to build multicriteria value measurement models [9,10,36], all of them requiring value judgements. We propose to use the Measuring Attractiveness by a Categorical Based Evaluation Technique (MACBETH), which asks only for qualitative pairwise comparison judgements of the difference in value between stimuli. A recent straightforward presentation of MACBETH can be found in [37]. MACBETH has been extensively applied in various evaluation contexts – see the references and mathematical foundations in [38] – namely to build reusable evaluation models [39], precisely the type of model required for faculty evaluation. The interactive application of MACBETH is visually supported by the M-MACBETH software [40].

3.2. Additive value measurement and optimisation procedures

The multicriteria additive value model constructed at IST has a two-level hierarchical structure detailed in Section 4.1, with the areas of activity j ($j=1,\ldots,M$) at the first level and the evaluation criteria i_j ($i_j=1,\ldots,N_j$) at a second level assumed to be exhaustive and nonredundant (see Section 4.1). Let P_{ij} be the descriptor of performance associated with each evaluation criterion i_j that belongs to the j area of activity (see Section 4.2). Let $P_{i_j}^d$ be the performance of faculty member d on criterion i_j, and $V_{i_j}^d = V_{i_j}(P_{i_j}^d)$ the partial value score obtained by faculty member d on criterion i_j, resulting from converting its performance into value through the use of the value function V_{ij} (see Section 4.5). Let w_{ij} be the weight given to criterion i_j (see Section 4.6). The area value score V_j^d for faculty member d taking only the criteria of area j into consideration is given by

$$V_j^d = \sum_{i_j=1}^{N_j} V_{i_j}^d w_{i_j} \tag{1}$$

with $\sum_{i_j=1}^{N_j} w_{i_j} = 1$ and $w_{i_j} > 0, \forall i,j$, and $V_{i_j}^d = 100$ when the performance of faculty member d in the evaluation criterion i_j equals the criterion target (see Section 4.3) set for the evaluation period under analysis and $V_{i_j}^d = 0$ when d has developed no activity related to that criterion in the same period. $(V_1^d,\ldots, V_j^d,\ldots, V_M^d)$ is the multicriteria value profile of faculty member d at the top level of the areas of activity.

Contrary to the criteria weights within each area, there is no fixed weight w_j assigned to each area of activity j ($j=1,\ldots,M$) and the area weights are free to vary within adequate intervals defined by lower and upper bounds $\underline{w_j}$ and $\overline{w_j}$ ($j=1,\ldots,M$), respectively. Therefore, the overall score $\overline{V}^d$ for faculty member d taking all the evaluation criteria into consideration will be given by solving the linear programming model (2), which guarantees for each faculty member d the maximum overall value score that can be attained with the interval weights defined and for her or his multicriteria area profile:

$$V^d = \max \sum_{j=1}^{M} V_j^d w_j \tag{2}$$

subject to $\underline{w_j} \le w_j \le \overline{w_j}$ and $\sum_{j=1}^{M} w_j = 1$ and $w_j \ge 0$.

Different model structures, either additive or multiplicative, can be used to capture both the quantity and quality components of faculty performance on each evaluation criterion [10,41]. In the structure of the IST model, for each evaluation criterion a bottom non-additive third level accounting for the quantitative and qualitative dimensions of academic activity was defined (see Fig. 3). The corresponding analytical structure in use is explained in detail in Section 4.2.

3.3. Assigning faculty members to rating categories

According to Portuguese law, every faculty member needs to be assigned to one, and only one, of a set of (ordered) rating categories of value. In the case of TUL, four categories – "excellent", "relevant", "sufficient" and "inadequate" – were pre-defined [42] to be adopted by all of its schools, including IST, which are free to define their own assignment procedure. The most common way is to define the categories by bounds of overall value and make the assignment based on preference relations [43], as in IST model (3), where V_1, V_2 and V_3 are the category bounds:

$$\textit{Rating category}^d = \begin{cases} \text{'Inadequate'}, & 0 \le V^d < V_1 \\ \text{'Sufficient'}, & V_1 \le V^d < V_2 \\ \text{'Relevant'}, & V_2 \le V^d < V_3 \\ \text{'Excellent'}, & V^d \ge V_3 \end{cases} \tag{3}$$

The bounds are the overall values of reference performance profiles (of hypothetical faculty members) that should be identified by the DM, for instance, with the support of the technique proposed in [44].

Conditional rules can also be used to sort faculty members, alternatively or in combination with a compensatory model of the type of model (3). For example, the DM might restrict the assignment to the "excellent" category to faculty members with a high overall value ($V^d \ge V_3$) and a number of scientific publications above the target, and the assignment to the "sufficient" category to faculty members with an overall score ranging from V_1 to V_2 and with a minimum involvement in all activities.

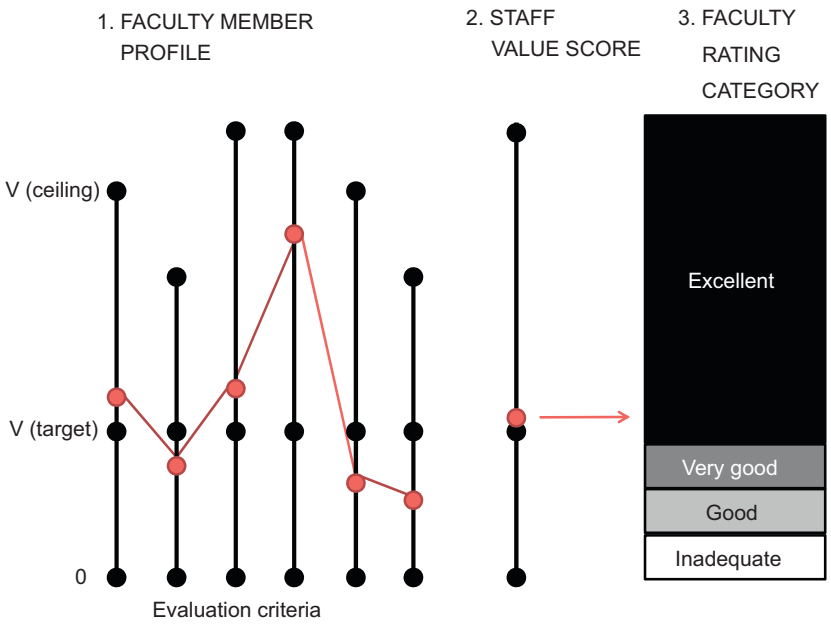

Fig. 2. Outputs from the proposed faculty evaluation model.

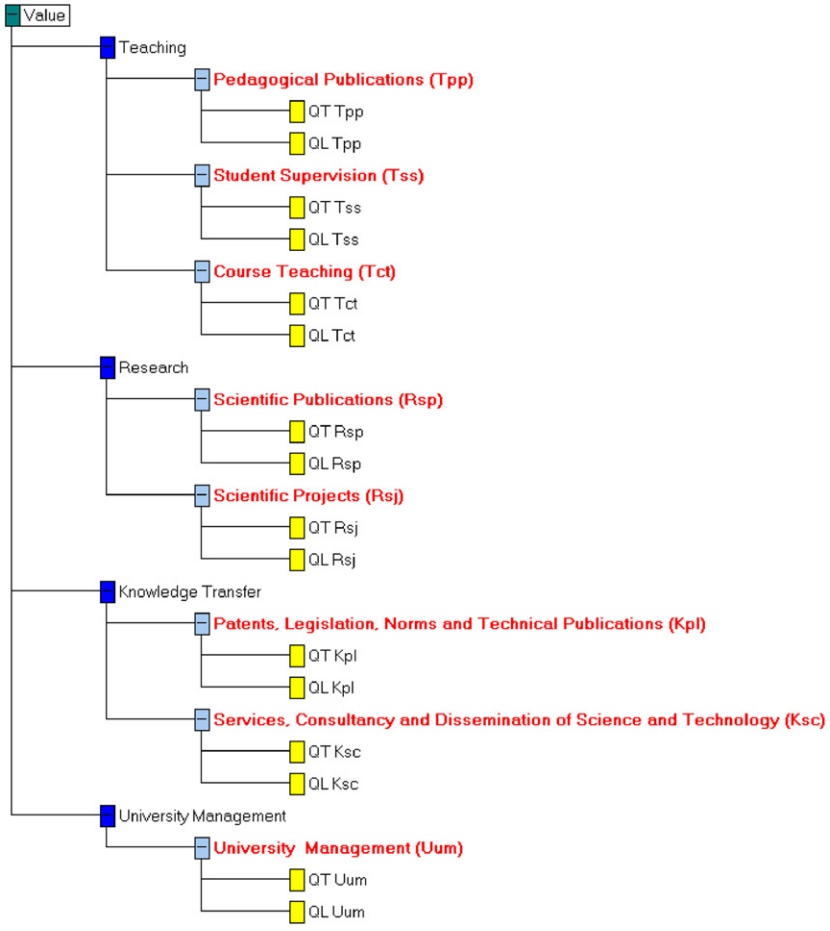

Fig. 3. Value tree.

3.4. *Model requisiteness*

While developing the multicriteria model, differences on individual views within the WG were debated in the decision conferences. The facilitator started by inviting the two participants with the most contrasting opinions (for example, the highest and lowest levels for a target) to explain their reasons, therefore motivating discussion within the WG, which usually resulted in revisions of initial views and convergence after a couple of rounds. In a few cases where a compromise was not

reached through discussion, a voting scheme was called upon and minority views were recorded for later sensitivity analysis to examine the extent of their effects on model outputs. As already observed by Phillips and Bana e Costa [34], "extensive sensitivity analyses show that many disagreements or uncertainties in the data make no difference to the overall results, and gradually a sense of common purpose emerges from the group".

From this process resulted an agreement within the WG on the proposed faculty evaluation model, which not only produces an overall value for each faculty member, but also generates key information for university management, namely, it allows for defining the multicriteria value profile of each faculty member (Fig. 2); comparing the performance of academic staff with performance targets; defining the multicriteria value profile of each faculty member at the top level of the evaluation areas; computing an overall value score for each faculty member, through the use of the optimisation procedure; and assigning faculty members to rating categories. By comparing faculty members' profile, overall scores and rating categories, the DM should analyse whether the model is adequate for evaluating faculty staff or whether it still needs to be revised (until it is "requisite").

The multicriteria model is flexible for application in the different scientific areas of the university, with some components of the model being common to all scientific areas, such as the set of evaluation criteria, and others varying according to the scientific area, such as targets and ceilings. Although the definition of targets and ceilings for each scientific area requires additional work, it may be necessary to ensure the requisiteness of the model across areas.

The next section presents selected features of the multicriteria model adopted by IST, as well as illustrates how multicriteria decision-aiding tools were used in building the model.

4. Multicriteria approach at IST

4.1. Value tree

The hierarchical structure of the model adopted by IST is depicted in the value tree of Fig. 3. There are four areas of activity on the first level, defined by law as teaching, research, knowledge transfer and university management [6]. The respective evaluation criteria were specified by the WG and appeared on the second level. For example, there are two research criteria: scientific publications and scientific projects (Rsp and Rsj). Each criterion includes both a quantitative component (QT) and a qualitative component (QL) that ensure exhaustiveness in evaluation. Indeed, when later confronted with the performance profiles of two hypothetical faculty members, *X* and *Y*, supposedly equivalent on all evaluation criteria, the WG could not identify any significant new evaluation aspect, either quantitative or qualitative, that might be invoked to differentiate *X* from *Y*. In structuring the criteria, overlaps detected between evaluation aspects were eliminated, within or across criteria, within or across areas, giving rise to a nonredundant set of criteria. At the end, the model was re-checked for double counting and the WG considered that *X* becoming better than *Y* on any evaluation component, either quantitative or qualitative, would always turn *X* globally better than *Y*. Therefore the value tree was considered concise and complete [45].

4.2. Descriptors of performance

4.2.1. Combining quantity and quality

Descriptors of performance account for the quantitative and qualitative dimensions of academic activity. For instance, evaluation related with scientific publications should consider both their number and quality (QT Rsp and QL Rsp, respectively, in Fig. 3). The performance $P_{i_j}^d$ of a faculty d in each evaluation criterion i_j is given by

$$P_{i_j}^d = QT_{i_j}^d f(QL_{i_j}^d) \tag{4}$$

where $QT_{i_j}^d$ and $QL_{i_j}^d$ are the quantitative and qualitative performance of d on i_j, respectively; $f(QL_{i_j}^d)$ is the value score for $QL_{i_j}^d$.

The multiplicative relationship between quantity and quality defined in Eq. (4) has implicit the ideas that one cannot assess quality if nothing has been produced, but quality is independent of the quantity produced while quantity is value dependent on quality. Simply said, the added value of publishing one more article is not independent of its quality. Fig. 4 helped discuss these assumptions with the WG, which agreed that, for publications of the same type, the difference in value between two articles of high quality and one article of high quality is Δ (graph 1), while the difference in value between two articles of low quality and one article of low quality is Δ' (graph 2), and with Δ' smaller than Δ; and the difference in value between two articles of high quality in comparison with two articles of low quality is Δ'' (graph 3), while the difference in value between one article of high quality and one article of low quality is Δ''' (graph 4), and with Δ''' equal to Δ''. Similar judgments helped generalise the multiplicative working hypothesis underlying Eq. (4) (see [41] for the technical assumptions used in multiplicative structures).

4.2.2. Describing quantitative performance

There are differences in the information used to compute quantitative and qualitative performance. Quantitative performance is measured using objective indicators validated in the literature [46] that can be computed automatically and do not demand specific intervention from the evaluator [11]. Accordingly, for each evaluation criterion, the WG has constructed an index that combines several indicators of quantitative performance. We present here the index that describes the quantity component of the scientific publications criterion:

$$QT_{Scientific\ \ Publications_{Research}} = \sum_{k=1}^{K} \frac{1}{Z_k}\left(T_k + \frac{1}{\rho}R_k\right) \tag{5}$$

with k being a scientific (and international) publication in the evaluation period ($k=1,\ldots,K$); T_k being the number of equivalent units for the type of publication k; R_k being the number of citations of k (excluding self-citations); ρ being the trade-off rate between citations and publications in the scientific area, defined on the basis of the average number of citations per publication [47] and Z_k being a correction factor for the number of authors of the publication (in the IST model, Z_k was not set equal to the number of authors of k, but rather as a function of it that establishes a compromise between creating incentives for joint research and rewarding each author's effort in publishing—see for details [42]).

The definition of T_k in Eq. (5) required trade-off judgements, from the WG, to convert one unit of each type of publication in units of a chosen reference type. The WG differentiated six types and ranked them in decreasing order of relative attractiveness as follows: one international book, one article in a type A (top) journal, one article in a type B journal, one chapter in an international book, one article in a type C journal, and one paper in international conference proceedings. Then, using MACBETH, the WG pairwise compared these different publication types in terms of their difference in attractiveness (including "nothing" published), giving rise to the consistent set of qualitative judgements displayed in the matrix of Fig. 5: for example, the judgement "extreme" in the cell highlighted in the matrix means that the WG judged publishing one article in a type A journal extremely more attractive than publishing one international book chapter. From these judgements, the

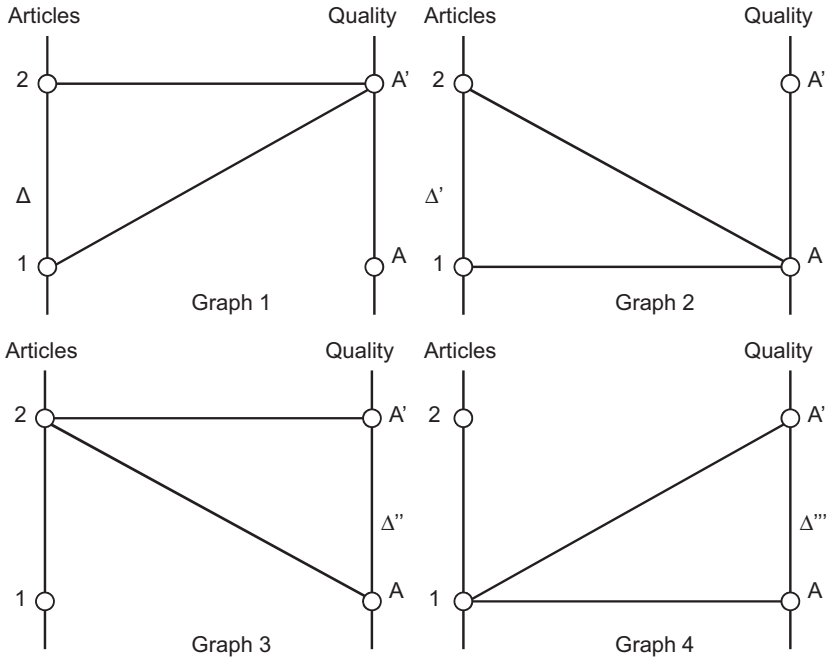

Fig. 4. Results from testing the relationship between quantity and quality.

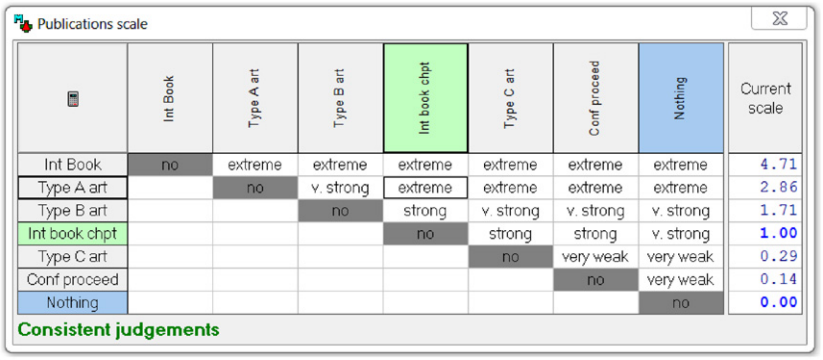

Fig. 5. MACBETH judgements for different types of scientific publications.

M-MACBETH software derived the scoring scale shown at the right of the matrix in Fig. 5, where one international book chapter was arbitrarily chosen as a reference publication and assigned a value score of 1. This basic MACBETH scale was afterwards subject to discussion and adjustment by the WG, until an agreement was reached on the final conversion scale (T) shown in Table 1 (in which the edition of one international book was later added as equivalent in value to one chapter in an international book). One can observe in Table 1 that, for instance, one article published in a type A journal is three times the contribution of one international book chapter. This may be a reasonable statement even for a significant number of publications. However, as well noted by Keeney et al. [26], one cannot interpret that publishing articles in type A journals is three times more important than publishing international book chapters.

Table 1
Equivalent units for different types of scientific publications.

Type of publication	T_k
International book	5.5
Article published in a type A journal	3
Article published in a type B journal	1.75
Chapter in an international book or edition of international book	1.0
Article published in a type C journal	0.3
Article published in conference proceedings	0.2

4.2.3. Describing qualitative performance

According to the principle of "controlled subjectivity" proposed by Arreola [11], the description of qualitative performance should be clear and make use of scales and methodological choices indicated by the literature [11,12]. Following these guidelines, a common descriptor of qualitative performance (expected to promote consistency across measurements) was included in the IST model. It is a five-level scale (see Table 2) constructed on the basis of the identification, for each criterion, of faculty members' pros and cons. These may be quality, behavioural, human, strategic or other aspects, as in peer review [11]. When pros and cons are balanced – or do not exist – the qualitative performance will not affect the quantitative performance (neutral effect in Table 2). On the other hand, if pros and cons are unbalanced, the quantitative performance should be rewarded or penalised, weakly or strongly, depending on the existence of "determinant" [48] characteristics, that is, pros and cons justifiably identified as "determinant", in the sense that they may strongly affect quantitative performance (strongly rewarding or penalising it, respectively).

Table 2
Descriptor of qualitative performance and effect value scale.

Levels of performance		$f(QL_{i_j})$
Strongly rewarding	*There is at least one determinant pro and no determinant con.*	1.5
Weakly rewarding	*The pros more than compensate the cons, and no pro and con is determinant.*	1.25
Neutral effect	*Pros and cons are balanced or do not exist.*	1
Weakly penalising	*The cons more than compensate the pros, and no pro and con is determinant.*	0.75
Strongly penalising	*There is at least one determinant con and no determinant pro.*	0.5

Table 3
Examples of performance targets for a 3-year period and for the evaluation criteria within the research and teaching areas of activity.

Evaluation criteria	Performance targets	Examples of targets (to be interpreted with $QL=1$)
Pedagogical publications (Tpp)	1.5	1 book chapter and 1 pedagogical text
Student supervision (Tss)	6	Supervision of 2 MSc theses per year, in a period of 3 years
Course teaching (Tct)	9	9 h of teaching per week with normal evaluation by students
Scientific publications (Rsp)	8	2 articles in type A journals and 2 chapters in international books (on the basis of no citations and no co-authors)
Scientific projects (Rsj)	1	Responsibility for 1 national R&D project

Table 4
Ceilings defined in terms of value for the evaluation criteria within the research and teaching areas of activity.

Evaluation criteria	Pedagogical publications (Tpp)	Student supervision (Tss)	Course teaching (Tcp)	Scientific publications (Rsp)	Scientific projects (Rsj)
V (ceiling)	500	300	300	600	500

According to Eq. (4), rewarding or penalising quantitative performance corresponds to multiplying it by a positive value score, which should be greater (resp. smaller) than 1 for rewarding (resp. penalising) qualitative levels (and, of course, equal to 1 when the effect of quality is neutral). The effect value scale shown in the last column of Table 2 resulted from MACBETH judgements of the WG, who fixed that the ratio between rewarding and penalising scores should not exceed 3. Take, for example, the scientific publications criterion. *X* and *Y* are two hypothetical faculty members having, in the evaluation period, publication portfolios with the same quantitative performance score. *X* is the single author of one article in a type A journal and *Y* is the single author of two chapters in international books and five papers in conference proceedings, and, to make it simple, *X* and *Y* have no co-authors and no citations in the period. The corresponding scores are $1\times3=3$ for *X* and $2\times1+5\times0.2=3$ for *Y* (see Table 1 and Eq. (5)). Suppose now that, in terms of quality, *X*'s evaluator identified one "determinant" pro: *X*'s article won a competition for outstanding contribution promoted by a highly reputed scientific society. On the contrary, suppose that the seven publications of *Y* are similar variants of another paper of *Y* published in the previous period in a type C journal—a repetition that *Y*'s evaluator could justifiably identify as a "determinant" con. If so, although *X* and *Y* have got the same quantitative performance score, the final performance score of *X* would be three times the final performance score of *Y*, respectively, $1.5\times3=4.5$ and $0.5\times3=1.5$ (see Table 2 and Eq. (4)).

4.3. Targets

The definition of performance targets is a useful instrument widely recognised in the literature on strategy and human resources management [49]. The WG decided that targets should be defined on each criterion and also by scientific area because, for instance, the expected number of publications by scientific area structurally differs. Table 3 provides examples of targets for the criteria within the research and teaching areas of activity: the performance targets in the second column are obtained by applying the indexes of quantitative performance to the levels of activity portrayed in the last column. For example, the performance target of 8 is obtained by applying Eq. (5) to the production of two articles in type A journals (corresponding to $2\times3=6$ units from Table 1) and two chapters in international books (corresponding to $2\times1=2$ units).

The target and the "zero" (null activity), to which correspond the value scores of 0 and 100 in Eq. (1), are the two reference levels later used in weighting the criteria (see Section 4.6).

4.4. Ceilings

It may be prudent to impose limits to the compensatory nature of a multicriteria model in the context of faculty evaluation. Indeed, if no limits exist, a faculty member might decide only to carry out one type of academic activity, disregarding any other activities, which might be undesirable. Thus the multicriteria model adopted by IST includes a ceiling for the value score attainable in each evaluation criterion, implying that there is a corresponding performance ceiling after which further activity by the faculty member adds little or no value to the school. Table 4 provides examples of value ceilings for the criteria within the research and teaching areas. They were related with the corresponding performance ceilings by assuming a linear value function. The requisiteness of the ceilings was later tested by asking the WG whether the rating category to which the model assigns a

faculty member who attains the ceiling in one criterion and has zero performance in all the other criteria is adequate.

4.5. Value functions

Value functions convert faculty performance into value for the school and its definition should account that different formats promote different incentives for the faculty members. For example, the linear value function, assumed when defining the ceilings, implies that the production of one more paper is always equally rewarded. However, after discussion, the WG agreed that it would be more realistic to adopt for all the criteria an S-shaped value function with the format presented in Fig. 6 (with a specific shape defined for each evaluation criterion). This function tends to reward performance close to the target: in the case of performance below the target, marginal increases in performance are gradually more valued; in the case of performance above the target, marginal increases in performance are valued at a decreasing marginal rate. As a result, the value function signals the relevance of targets.

Eqs. (6a)–(6c) define the S-shaped value function by two exponential functions, corresponding to two branches (below and above the target) that respect the "delta property", reflecting a "constant trade-off attitude" [10] that the WG considered to be a desirable property in the context of faculty evaluation. For an additive value function and for an attribute X_i, a value function verifies the delta property when it is true that x_i^m is the mid-value performance of $[x_i', x_i'']$, then $x_i^m + \delta$ is the mid-value performance of $[x_i' + \delta, x_i'' + \delta]$, for any δ (the proof and the rationale for the constant trade-off condition are discussed in [10]):

$$V_{i_j}^d = \left[1 - \exp\left(-P_{i_j}^d/\rho_1\right)\right] \Big/ \left[1 - \exp\left(-P_{target_{i_j}}/\rho_1\right)\right], \quad \text{if } 0 \le P_{i_j}^d \le P_{target_{i_j}} \tag{6a}$$

$$V_{i_j}^d = \left[1 - \exp\left(-\frac{P_{i_j}^d - P_{target_{i_j}}}{\rho_2}\right)\right] \Big/ \left[1 - \exp\left(-\frac{P_{ceiling_{i_j}} - P_{target_{i_j}}}{\rho_2}\right)\right], \quad \text{if } P_{target_{i_j}} \le P_{i_j}^d < P_{ceiling_{i_j}} \tag{6b}$$

$$V_{i_j}^d = V_{ceiling_{i_j}}, \quad \text{if } P_{i_j}^d \ge P_{ceiling_{i_j}} \tag{6c}$$

with $P_{target_{i_j}}$ being the performance target for i_j; $P_{ceiling_{i_j}}$ and $V_{ceiling_{i_j}}$ being, respectively, the performance and value ceilings (defined as explained in the previous section) and ρ_1 and ρ_2 the parameters that determine the shape of the exponential branches and need to be selected for each evaluation criterion. ρ_1 is negative, and the lower the value, the closer the value function to a linear function; ρ_2 is positive, the higher the value, the closer the value function to a linear function.

To ensure that an S-shaped value function verifying the delta property was compatible with the intuitions of the WG, a test using MACBETH judgements was employed (other examples of tests to identify whether a value function is compatible with pairwise comparisons were earlier explored by Salminen et al. [50]). Fig. 7 displays an example of a test for the student supervision evaluation criterion, with performance varying between 0 and a ceiling of 18 theses supervised and with a target of 6 theses (see the interpretation of the target and the value ceiling in Tables 3 and 4, respectively). To apply the test, firstly, qualitative judgements for the added value of one extra thesis are made, comparing consecutive numbers of theses and filling the matrix of judgements. Secondly, a set of constraints that represent conditions of constant trade-off attitude is introduced into the M-MACBETH software. Five of those constraints are illustrated in Fig. 7: the first three constraints define the constant trade-off for a number of publications above the target, while the next two constraints define similar conditions for numbers below

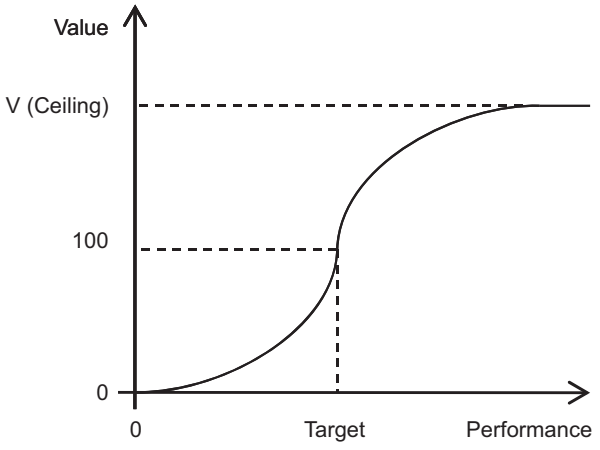

Fig. 6. S-shaped value function.

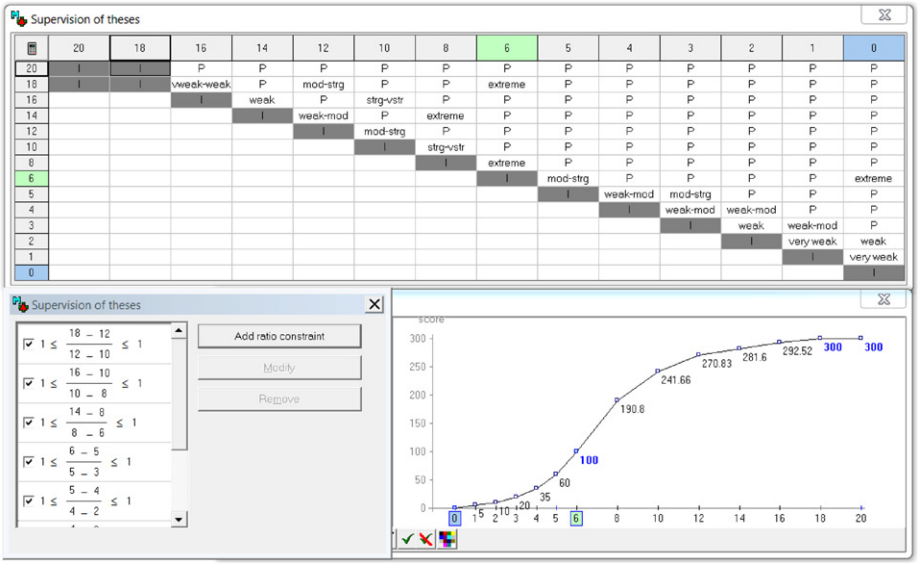

Fig. 7. Testing whether preferences respect the constant trade-off attitude condition.

the target. The first three constraints read as 8 is the mid-value performance between 6 and 14 theses supervised and, adding 2 to both bounds, 10 is the mid-value performance between 8 and 16 (and similarly, adding 2 to these bounds, 12 is the mid-value performance between 10 and 18). Thirdly, M-MACBETH is used to analyse whether there is a value function that is compatible with the matrix of judgements and with the set of constraints. The value function displayed in Fig. 7 is compatible with that information, providing a piecewise linear function that is an approximation of an S-shaped value function with two exponential branches defined below and above the target.

4.6. Weighting criteria

As already highlighted in Section 2, the use of inappropriate weighting procedures is undoubtedly a common problem in evaluation studies [26] and it is critical in the framework of an additive value model [28,48]. Because, the weights are scaling constants allowing the additive aggregation of value scores on different criteria and, consequently, their assessment requires value trade-offs [41]. Given the peculiar structure of the additive model for faculty evaluation proposed in Section 3.2 (see Eqs. (1) and (2) and Fig. 3), the weighting process, supported by the M-MACBETH software, was conducted in such a way that its final output consists of fixed weights for the evaluation criteria within each area of activity (Eq. (1)) and interval weights across the areas (Eq. (2)).

Firstly, the WG was introduced to the concept of swing weighting [9] and asked to rank, in decreasing order of relative importance, the swings from null performance to the target on the eight evaluation criteria. Appropriate examples were discussed to show that rank reversals could naturally occur when the target performances were varied. For instance, referring to the targets in Table 3, a swing from 0 to 9 h of teaching per week was judged by the WG to be more attractive than a swing from 0 to 6 MSc theses supervised during 3 years, but a decrease of 3 h in the teaching target would make, the WG said, the new 0–6 h teaching swing indifferent to the unchanged supervision swing. Then, MACBETH judgements were assessed from the WG for each swing and for each pair of swings, giving rise to the consistent matrix of MACBETH weighting judgements of Fig. 8. For example, although the swings of 0 – 9 h [Tct] and 0 – 6 MSc theses [Tss] were both considered to be moderately important, the difference between the former and the latter was also judged to be moderate. To guarantee a shared understanding of the weighting questioning process, the WG was asked to validate the following judgemental statement, which is equivalent to the above interpretation in terms of swings in performance: the difference in overall value between a faculty member who has achieved the 9 h target performance in teaching but has not carried out any other activity [Tss] and another faculty member who has achieved the 6 MSc theses target for student supervision but having no other activity [Tct] is moderate.

The histogram in Fig. 8 shows the respective MACBETH weights (in percentages), the requisiteness of which was discussed and gave rise to the revision of some input judgments, for which MACBETH generated new weights, and so on, in a recursive, iterative and interactive process that ended when the WG 'gut feelings' and intuitions matched the final criteria weights shown in the last row of Table 5. To facilitate this process in such a way that the WG could easily validate proportions between weights of criteria belonging to the same area, the weights in percentages were rescaled so that the less important swing (actually, two swings) was worth 1. This promptly revealed that, for instance, the most important swing (from 0 to the target in scientific publications) was agreed to value 6 times more than swinging from 0 to the target in pedagogical publications.

Table 5 also shows the derived area weights, which constituted the basis for the definition of lower and upper bounds for area weights. For this purpose, those derived area weights were firstly rescaled so that the highest of them valued 100, resulting in the swing-derived area weights shown in the second column of Table 6. The next two columns in Table 6 present the minimum and maximum acceptable swings defined by the WG. For example, the WG agreed that swinging from no activity to the targets simultaneously in all the three criteria within the teaching area would worth between 50% and 90% of swinging from no activity to the targets simultaneously in the criteria of the research area. One can now apply Eqs. (7) and (8) to the acceptable limits in order to find lower and upper bounds for the intervals of variation of the area weights (see the two last columns of Table 6). These intervals were finally adjusted by the WG to the final intervals shown in the second row of Table 7. Contrary to the area weights, the weights of the evaluation criteria were taken as fixed within each area (see the last row of Table 7):

$$\overline{w_j} = \frac{s^{max_j}}{s^{max_j} + \sum_{y \neq j} s^{min_y}} \tag{7}$$

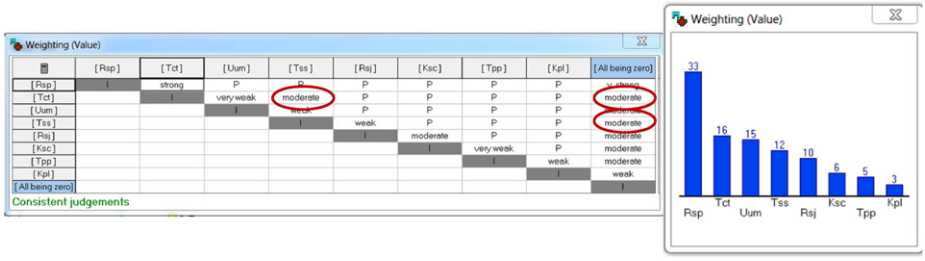

Fig. 8. Example of a set of weights obtained using the MACBETH approach (names of evaluation criteria in Fig. 3).

Table 5
Area and criteria weights (names of evaluation criteria in Fig. 3).

Areas of activity	**Teaching**			**Research**		**University management**	**Knowledge transfer**	
Area weights (w_j)	6 (30%)			8 (40%)		3 (15%)	3 (15%)	
Evaluation criteria	Tct	Tss	Tpp	Rsp	Rsj	Uum	Ksc	Kpl
Criteria weights (w_{ij})	3 (15%)	2 (10%)	1 (5%)	6 (30%)	2 (10%)	3 (15%)	2 (10%)	1 (5%)

Table 6
Generating intervals of variation of the area weights.

Areas of activity	Swing-derived area weights	Minimum acceptable	Maximum acceptable	Lower bound (%)	Upper bound (%)
Teaching	75	50	90	20.8	42.9
Research	100		100 (reference)	35.7	58.8
Knowledge transfer	37.5	10	50	4.2	23.8
University management	37.5	10	40	4.0	20.0

Table 7
Criteria weights and interval area weights (names of evaluation criteria in Fig. 3).

Areas of activity	Teaching			Research		University management	Knowledge transfer	
Area weights (w_j)	20–40%			40–60%		0–20%	0–20%	
Evaluation criteria	Tct	Tss	Tpp	Rsp	Rsj	Uum	Ksc	Kpl
Criteria weights (w_{ij})	3/6	2/6	1/6	6/8	2/8	1/1	2/3	1/3

$$\underline{w_j} = \frac{s^{min_j}}{s^{min_j} + \sum_{y \neq j} s^{max_y}} \quad (8)$$

with s_j^{min} being the minimum swing for j and s_j^{max} the maximum swing for j.

In the last step, intervals of weights were built for each area of activity. The second column in Table 6 presents the (point) swings used in Table 5, after these had been normalised so that the highest weight is 100 (the reference is built against the research area of activity).

5. Discussion

Following the revision in 2009 of the statute that defines the careers of faculty members [6], each school in Portugal is now required to adopt a faculty evaluation model that should provide information on how faculty members can improve their performance and adequately distinguish between levels of faculty performance, while accounting for specificities across scientific areas. IST managers recognised this moment as a unique opportunity to communicate to faculty staff the academic activities judged to be strategically important to the school.

Two aspects regarding model building and using the faculty evaluation model deserve special attention. Firstly, although the model was tested for two types of redundancies – within criteria and across criteria redundancies – there might be cases that require special analysis. In order to avoid within criteria redundancies, using the model should account for potential overlaps between quantitative and qualitative performance. Recall the descriptor of qualitative performance and the effect value scale in Table 2. Suppose that a faculty member is, for example, the author of an article that has been cited extensively, contributing significantly to the author's quantitative performance. Suppose also that the evaluator considered the high impact of that article as a "determinant" pro and consequently the evaluator rewarded the author with the highest effect value (academic studies indicate that citations are a good indicator of the impact of publications, and citations are correlated with the evaluation of the content of publications by peers [51,52]). In this situation, there exists double counting in quantitative and qualitative performance. In other cases, there might be potential overlaps across criteria. For instance, consider an international textbook that may also be viewed as contributing for the scientific development of an area, for example the book by Kirkwood [10]. Should it be classified as a pedagogical or as a scientific publication (within the teaching and research areas, respectively) or both? Secondly, an effect value scale sets the balance between the objective and subjective components of the model criteria (i.e., components not depending on the evaluator and components based upon peer review by the evaluator, respectively). The choice at IST of an effect value scale ranging between 0.5 and 1.5 (see Table 2) derives from the desire that the ratio between rewarding and penalising scores should not exceed 3. In cases where the DM wants to privilege the peer review component, it would be enough to fix a higher ratio. On the contrary, a ratio of 1 would model the extreme preference for a pure objective evaluation.

Before submitting the multicriteria model to the DM (the Management Board of IST) the WG analysed its outputs for a wide range of faculty profiles and performed sensitivity analyses on different model parameters, namely, on those ones that had originated some disagreement within the WG. Some minor adjustments gave rise to a requisite model at the eyes of the WG. Then the DM has opened the discussion to the faculty and made some decisions following feedback and suggestions from the school members. One important decision of the DM concerned value functions. Despite acknowledging the adequacy of the proposed S-shaped function, the DM has chosen, on grounds of simplicity and for political reasons, to adopt a linear function truncated for the value of the ceiling, as exemplified in Fig. 9 for the scientific publications criterion (one more paper published values always the same, for publications of the same type and below the ceiling). Linear value functions were realistically seen as less subject to controversy within the school than S-shaped ones.

The final multicriteria model was adopted by IST in 2010 to evaluate faculty members for their activities in the 2004–2007 and 2008–2009 periods. Since these evaluations were made retrospectively, each faculty member could choose between being evaluated using the faculty evaluation model or through holistic evaluation of their CV. The vast majority opted for the model evaluation. A spreadsheet implementation of the model was produced and made available for each faculty member to fill in with her or his data in each of the two periods and for the respective evaluator to validate and complete with qualitative data about the assessed. Moreover, as the model will also be used in the evaluation period underway (2010–2012), faculty can easily monitor the value of their production using the spreadsheet programme (the interested reader can request a copy to the IST Management Board via the corresponding author).

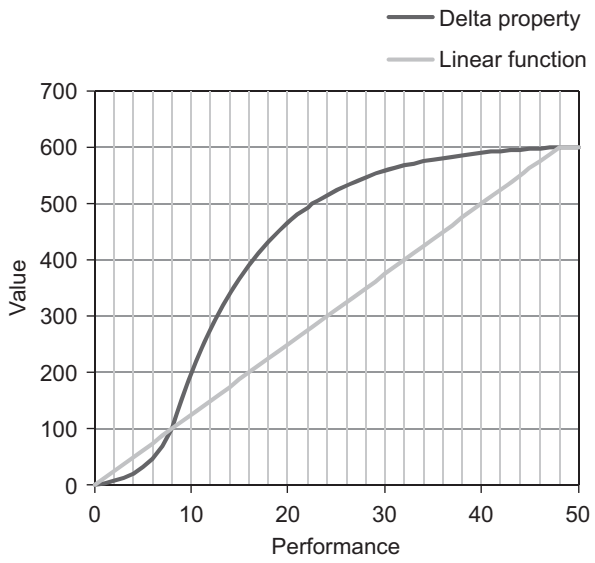

Fig. 9. Value function chosen by IST vs. an alternative S-shaped value function (for scientific publications).

Notwithstanding the sound theoretical foundations of the proposed multicriteria model, its effectiveness is yet to be confirmed by practical application on a large scale. First, despite the fact that it was designed to be applied to different scientific domains, only the implementation of the model and further analysis will show whether it deals effectively with differences across scientific domains and whether adjustments are required. Second, there are concerns about the calibration of the model when different targets and other values across scientific domains are used. Finally, the incentives motivated by the model adoption have not been studied in detail.

The proposed model might be improved in several ways. At first, it should be better informed by literature in specific areas, such as on teaching indicators and on research indicators. For example, should the evaluation of teaching be performed only by students, or should peers also review the content of teaching? If peer reviewing of the context of teaching is required, which methods for evaluation are available and have been validated? Which is the best indicator of the impact of publications on the community? Moreover, the model was developed without detailed information on the performance of IST academic staff for many indicators. The use of high-quality data and of scientific metrics might contribute to building a sounder model and a higher level of acceptation. Participatory mechanisms to improve the model should be developed, so that its adoption fosters an evaluation culture. The implementation of the model requires the collection of a wide set of information about staff activity and generates a wide range of information on the school performance. There is scope for developing multiple criteria interactive analysis tools that might help in the collection and analysis of model inputs and outputs. Some of the multicriteria methods in use can be further developed. For example, which procedures should be used for validating the descriptors of performance in use? Should category bounds be linked with targets and with ceilings? Last but not the least, the proposed model should be tested within schools other than engineering ones to investigate whether it respects the evaluation context in other scientific domains. In fact, one should be aware that disciplines might differ in their preferred approach to knowledge transmission.

One final note to refer that the socio-technical process developed at IST and addressed in this article followed a modern [53] decision analysis approach that is a mixture of "engineering science" and "clinical art", as defined by Buede [54], where we acted as impartial facilitators and decision analysts. We believe that, however, it is not contradictory with this role to assert that, as faculty members, we subscribe the 'Criteria for Accreditation' defined by the Southern Association of Colleges and Schools in the USA [18]: "An institution must conduct periodic evaluations of the performance of individual faculty members. The evaluation must include a statement of the criteria against which the performance of each faculty member will be measured. The criteria must be consistent with the purpose and goals of the institution and be made known to all concerned. The institution must demonstrate that it uses the results of this evaluation for the improvement of [its] faculty and its educational program".

Acknowledgements

The authors gratefully acknowledge that this paper is based on work developed at the Instituto Superior Técnico, which they would like to thank for the opportunity given to publish the case, and would also like to thank all the colleagues who contributed to the development of the IST model building process, namely the members of the Working Group. The authors also thank Amílcar Sernadas, Larry Phillips, Michael Cassidy, Paulo Martins and two anonymous referees for their thorough and insightful comments on earlier versions of this paper. The authors remain responsible for any omissions and inaccuracies.

References

[1] Etzkowitz H. Research groups as 'quasi-firms': the invention of the entrepreneurial university. Research Policy 2003;32:109–21.
[2] Coccia M. Measuring scientific performance of public research units for strategic change. Journal of Informetrics 2008:183–94.
[3] Huber MT. Faculty evaluation and the development of academic careers. New Directions for Institutional Research 2002;114:73–83.
[4] Murias P, Miguel JCd, Rodríguez D. A composite indicator for university quality assesment: the case of Spanish higher education system. Social Indicators Research 2008;89:29–146.
[5] Agencia Nacional de Evaluación de la Calidad y Acreditación. Programa de Evaluación de Professorado para la Contratación: Principios y Orientaciones para la Aplicación de los Criterios de Evaluación. Agencia Nacional de Evaluación de la Calidad y Acreditación; 2007.
[6] Ministério da Ciência Tecnologia e Ensino Superior. Decreto-Lei no. 205/2009. Diário da República. 2009;1.ª série.
[7] Elmore HW. Toward objectivity in faculty evaluation. Academe 2008;94:38–40.
[8] Lane J. Let's make science metrics more scientific. Nature 2010;464:488–9.
[9] von Winterfeldt D, Edwards W. Decision analysis and behavioral research. Cambridge University Press; 1986.
[10] Kirkwood CW. Strategic decision making: multiobjective decision analysis with spreadsheets. Belmont, California: Duxbury Press; 1997.
[11] Arreola RA. Developing a comprehensive faculty evaluation system: a guide to designing, building, and operating large-scale faculty evaluation systems, 3rd ed. Anker Publishing Company; 2007.
[12] National Academy of Engineering. Developing metrics for assessing engineering instruction: what gets measured is what gets improved: report from the steering committee for evaluating instructional scholarship in engineering. The National Academies Press; 2009. p. 52.
[13] Mills M, Hyle AE. Faculty evaluation: a prickly pair. Higher Education. 1999;38:351–71.
[14] Adler NJ, Harzing AW. When knowledge wins: transcending the sense and nonsense of academic rankings. Academy of Management Learning and Education 2009;8:72–95.
[15] Grant J, Fogarty T. Faculty evaluation as a social dilemma: a game theoretic approach. Accounting Education 1998;7:225–48.
[16] Kao C, Pao H-L. An evaluation of research performance in management of 168 Taiwan universities. Scientometrics 2009;78:261–77.
[17] Turban E, Zhou D, Ma J. A group decision support approach to evaluating journals. Information & Management 2004;42:31–44.
[18] Campion WJ, Mason DV, Erdman H. How faculty evaluations are used in Texas Community Colleges. Community College Journal of Research and Practice 2000;24:169–79.
[19] Desselle SP, Mattei TJ, Vanderveen RP. Identifying and weighting teaching and scholarship activities among faculty members. American Journal of Pharmaceutical Education 2004;68:1–11.

[20] Donovan C. The qualitative future of research evaluation. Science and Public Policy 2007;34:555–63.
[21] Wolansky WD. A multiple approach to faculty evaluation. Education 2001; 97:81–96.
[22] Politis Y, Siskos Y. Multicriteria methodology for the evaluation of a Greek engineering department. European Journal of Operational Research 2004; 156:223–40.
[23] Giménez VM, Martínez JL. Cost efficiency in the university: a departmental evaluation model. Economics of Education Review 2006;25:543–53.
[24] Uzoka F-ME. A fuzzy-enhanced multicriteria decision analysis model for evaluating university academics' research output. Information Knowledge Systems Management 2008;7:273–99.
[25] Mustafa A, Goh M. Multi-criterion models for higher education administration. OMEGA—The International Journal of Management Science 1996;24: 167–78.
[26] Keeney RL, See KE, von Winterfeldt D. Evaluating academic programs: with applications to U.S. graduate decision science programs. Operations Research 2006;54:813–28.
[27] Edwards W, Barron FH. SMARTS and SMARTER: improved simple methods for multiattribute utility measurement. Organizational Behavior and Human Decision Processes 1994;60:306–25.
[28] Keeney RL, Value-Focused Thinking A. Path to creative decisionmaking. Harvard University Press; 1992.
[29] Billaut J-C, Bouyssou D, Vincke P. Should you believe in the Shanghai ranking? An MCDM view Scientometrics 2010;84:237–63.
[30] Phillips L. Decision conferencing. In: Edwards W, Miles RF, von Winterfeldt D, editors. Advances in decision analysis: from foundations to applications. Cambridge University Press; 2007. p. 375–99.
[31] Phillips L. Decision conferencing/facilitated workshops. In: Melnick EL, Everitt BS, editors. Encyclopedia of quantitative risk analysis and assessment. Wiley; 2008. p. 451–9.
[32] Schein EH. Process consultation revisited: building the helping relationship. MA: Addison Wesley Longman; 1998.
[33] Bana e Costa CA, Ensslin L, Correa EC, Vansnick J-C. Decision support systems in action: integrated application in a multicriteria decision aid process. European Journal of Operational Research 1999;113:315–35.
[34] Phillips LD, Bana e Costa CA. Transparent prioritisation, budgeting and resource allocation with multi-criteria decision analysis and decision conferencing. Annals of Operations Research 2007;154:51–68.
[35] Phillips LD. A theory of requisite decision models. Acta Psychologica 1984;56: 29–48.
[36] Belton V, Stewart TJ. Multiple criteria decision analysis: an integrated approach. Springer; 2001.
[37] Bana e Costa CA, Vansnick J-C, De Corte J-M. MACBETH (measuring attractiveness by a categorical based evaluation technique). In: Cochran JJ, editor. Wiley encyclopedia in operational research and management science. New York: Wiley; 2011. p. 2945–50.
[38] Bana e Costa CA, De Corte J-M, Vansnick J-C. On the mathematical foundations of MACBETH. In: Figueira J, Greco S, Ehrgott M, editors. Multiple criteria decision analysis: the state of the art surveys. Springer; 2005. p. 409–42.
[39] Bana e Costa CA, Lourenço JC, Chagas MP, Bana e Costa JC. Development of reusable bid evaluation models for the Portuguese Electric Transmission Company. Decision Analysis 2008;5:22–42.
[40] Bana e Costa CA, De Corte J-M, Vansnick J-C. MACBETH. In: Department OR, editor. Working paper LSEOR 0356. London: London School of Economics of Political Science; 2003.
[41] Keeney RL, Raiffa H. Decisions with multiple objectives: preferences and value tradeoffs New York. Wiley; 1976.
[42] Universidade Técnica de Lisboa. Regulamento de Avaliação de Desempenho dos Docentes do Instituto Superior Técnico, 51. Diário da República; 2010.
[43] Öztürk M, Tsoukiàs A, Vincke P. Preference modelling. In: Figueira J, Greco S, Ehrgott M, editors. Multiple criteria decision analysis: state of the art surveys; 2005. p. 27–72.
[44] Bana e Costa CA, Oliveira RC. Assigning priorities for maintenance, repair and refurbishment in managing a municipal housing stock. European Journal of Operational Research 2002;138:380–91.
[45] Keeney RL. Developing objectives and attributes. In: Edwards W, Miles Jr. RF, von Winterfeldt D, editors. Advances in decision analysis: from foundations to applications. Cambridge: Cambridge University Press; 2007. p. 104–28.
[46] Meredith JR, Steward MD, Lewis BR. Knowledge dissemination in operations management: published perceptions versus academic reality. OMEGA—The International Journal of Management Science 2011;39:435–46.
[47] Althouse BM, West JD, Bergstrom CT, Bergstrom T. Differences in impact factor across fields and over time. Journal of the American Society for Information Science and Technology 2009;60:27–34.
[48] Bana e Costa CA, Corrêa EC, De Corte J-M, Vansnick J-C. Facilitating bid evaluation in public call for tenders: a socio-technical approach. OMEGA—The International Journal of Management Science 2002;30:227–42.
[49] Drucker P. Management by results. Collins 1993.
[50] Salminen P, Korhonen P, Wallenius J. Testing the form of a decision-maker's multiattribute value function based on pairwise preference information. Journal of the Operational Research Society 1989;40:299–302.
[51] Meho LI, Sonnenwald DH. Citation ranking versus peer evaluation of senior faculty research performance: a case study of Kurdish scholarship. Journal of the American Society for Information Science 2000;51:123–38.
[52] Donohue JM, Fox JB. A multi-method evaluation of journals in the decision and management sciences by US academics. OMEGA—The International Journal of Management Science 2000;28:17–36.
[53] Belton V. Multi-criteria problem structuring and analysis in a value theory framework. In: Gal T, Stewart TJ, Hanne T, editors. Multicriteria decision making: advances in MCDM models, algorithms, theory and applications. Springer; 1999. p. 12.1—12.32.
[54] Buede DM. Applied decision analysis: engineering science vs. clinical art. In: Proceedings of the international conference on cybernetics and society. Denver; 1979. p. 397–403.