

TITLE: A Multicriteria Decision Analysis Model for Faculty Evaluation

AUTHORS: Bana e Costa, C., Oliveira, M. D., Martins, P.A.F., Sernadas, A. (*Instituto Superior Técnico*, Technical University of Lisbon, Portugal)

ABSTRACT: The institutional and legal setting where universities operate has undergone major transformations in recent years. There is a global trend towards increasing social and financial accountability of universities and a general demand for higher transparency and responsibility in decisions related to academic staff performance appraisal. Bringing faculty evaluation systems in line with these changes became a priority in many countries. For example, in Portugal, all universities are presently working on regulations defining the evaluation systems to be applied to all academic staff. The implementation of a faculty evaluation system constitutes a challenge and an opportunity for strategically aligning the activity of the academic staff with the university goals.

However, despite a growing interest on performance appraisal of faculty, literature in the area does not provide comprehensive models for the evaluation of academic staff. The few reported studies that attempt to evaluate the overall activity of academics and the existing metrics do not capture the full range of academic activities. *Ad hoc* scoring systems lack theoretical soundness.

This article approaches faculty evaluation from an innovative comprehensive perspective. Based on concepts and methods of multiple criteria value measurement, it proposes a multicriteria 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. The model has a two-level hierarchical additive structure, and accounts for the quantitative and qualitative dimensions of academic activity at a bottom non-additive third level. The model uses appropriate methods for building value functions and assessing inter-criteria weighting information, and allows to compare performance of academic staff with performance targets reflecting strategic policy concerns of university management. The model allows to define the multicriteria value profile of each faculty member; and to compute his/her overall value through an optimization procedure that makes use of a flexible system of weights.

The model was adopted by *Instituto Superior Técnico* (IST), the engineering school of the Technical University of Lisbon, to be employed in decisions regarding recruitment, promotion, granting tenure, and recognizing excellence in academic activities.

KEYWORDS: multicriteria decision analysis, faculty evaluation, decision support systems

1. INTRODUCTION

In the past years there have been considerable changes in the university system of organization and funding. The traditional activities of teaching, research and service are increasingly committed to the needs of society (Etzkowitz, 2003), and universities have been assuming active responsibilities within the economy (Coccia, 2008).

In addition, the institutional and legal setting where many universities operate has undergone major key transformations. There is a global trend towards increasing social and financial accountability (Huber, 2002), as well as a general demand for higher transparency and responsibility in decisions related to the evaluation of faculty members.

Bringing faculty evaluation systems in line with the changes in the university system became a priority in many countries around the world. Faculty evaluations are 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 (Huber, 2002). In Europe, the need for developing evaluation tools is recognized both at the national level and at the EU supra-national level (Murias, Miguel, & Rodríguez, 2008). For example, in Spain, national rules have been defined in recent years for evaluation of academic staff (Agencia Nacional de Evaluación de la Calidad y Acreditación, 2007). In Portugal, all universities are presently defining faculty evaluation systems (Ministério da Ciência Tecnologia e Ensino Superior, 2009).

As a result of these developments, there is a challenge and an opportunity for each university to align the activity of their faculty members with its mission and strategic plans. Universities are expected to make decisions on recruiting, promoting, granting tenure and rewarding excellency based on objective evaluation criteria and supported by appropriate tools.

Despite the international growing interest on performance appraisal on university activities, and in particular of faculty evaluation, there is only a few number of studies that attempt to evaluate the overall activity of the academic staff (Elmore, 2008) and “existing metrics do not capture the full range of activities that support and transmit scientific ideas” (Lane, 2010) (page 488).

This paper results from what authors consider the key challenges in faculty evaluation, namely the need to develop comprehensive evaluation systems, based on methodologically sound procedures that can adequately reflect differences between the academic staff, taking into account the university mission, and applicable to all faculty members and scientific areas while respecting their specificities.

This article proposes an innovative comprehensive approach for faculty evaluation. Multiple criteria value measurement is the right framework for faculty evaluation because it is supported by strong theoretical foundations (see for example (Von Winterfeldt & Edwards, 1986)) and is capable of addressing the socio-technical and multidimensional nature of the evaluation problem, where multiple and conflicting viewpoints and objectives need to be taken into consideration. Moreover, multicriteria decision analysis is flexible enough to integrate both quantitative and qualitative information, in line with the recommendations and guidelines on how to build comprehensive faculty evaluation models (Arreola, 2007) (National Academy of Engineering, 2009). The proposed faculty evaluation model was designed within the legal and institutional context of the Portuguese universities, and to be used by *Instituto Superior Técnico* (IST), an engineering school with 778 faculty members working in a wide variety of scientific areas (ranging from mathematics, physics and chemistry to most branches of engineering, architecture and management). The paper is structured in four parts: the first part presents a brief state-of-the-art on faculty evaluation methods, the second part presents the features of the adopted multicriteria modelling approach, the third part describes the model developed at IST and the fourth part discusses what was achieved and what is still ahead.

2. LITERATURE REVIEW

Personnel management, self-improvement, growth and development of faculty members, and improvement of the quality of instruction in schools are understood to be key objectives for faculty evaluation (Mills & Hyle, 1999). Given the nature of academic activity and the organizational structure of universities, evaluation systems of academics in use in universities are mostly based on peer review. Nevertheless, differences exist on the information basis and methods that peers might use in the evaluation process. While several authors sustain that it is possible to measure with some precision faculty performance and that performance measurements might be used in university management (Arreola, 2007), others defend that scientific activities cannot be fully measured given current knowledge and available indicators, and that the use of measurement tools might affect researchers' autonomy and might lead to undesirable effects (Adler & Harzing, 2009). Different opinions are partly explained by methodological difficulties related with:

- It is hard to measure an individual faculty members' 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 (Grant & Fogarty, 1998). Differing values given to these activities are apparently neither appreciated nor systematically communicated (Grant & Fogarty, 1998). It is difficult to define which activities to include in scholarship (Huber, 2002) and to find appropriate indicators for performance measurement (Adler & Harzing, 2009). Evaluation methods are sensitive to selected indicators and to data sources (Kao & Pao, 2009).

- Faculty evaluation models typically make use of objective approaches and/or of subjective approaches (Turban, Zhou, & Ma, 2004). Objective approaches do not depend on the evaluator (for example, using citation counts) and might generate unintended results because of problems with data (such as with bibliometric data) (Lane, 2010), generating biases in evaluation (Coccia, 2008). Subjective approaches can be influenced by personal biases or by some lack or insufficient knowledge or experience by some group members (Turban et al., 2004). There has been little research on how to adequately integrate objective and subjective approaches (Turban et al., 2004).
- Given that faculty evaluation implicitly incorporates many beliefs about academic careers and institutional policy, generates different costs and shapes power relationships between stakeholders, as well as interacts with the balance between personal and departmental goals in academia (Grant & Fogarty, 1998), it is not an easy task to build and promote changes in evaluation systems.
- Faculty evaluation literature is spread across several areas. While some professions have had extensive discussions about evaluation models and tools (being the case of the pharmacy and accounting communities (Grant & Fogarty, 1998) (Desselle, Mattei, & Vanderveen, 2004)), there has been undervaluing or underreporting of research for some communities (ex: social sciences) (Donovan, 2007). Most evaluation studies explicitly state their area of applicability.

Although to date no movement has emerged to standardise the evaluation process and maximize objectivity while linking productivity in an empirical fashion to rewards (Elmore, 2008), multiple institutions have advocated the need to develop an evaluation culture in university systems (Murias et al., 2008) and to create more comprehensive evaluation systems. This is the case of the National Academy of Engineering in the US (National Academy of Engineering, 2009) and the director of the Science of Science & Innovation Policy programme from the National Science Foundation in the US (Lane, 2010).

Analysis of evaluation literature in the university context shows that most studies reported carried out comparative analysis of universities, faculties, departments or research units (such as (Wolansky,

2001) (Giménez & Martínez, 2006) (Coccia, 2008)), 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 the university production and a central 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 (Mills & Hyle, 1999) (Campion, Mason, & Erdman, 2000) (Desselle et al., 2004). Some propose conceptual frameworks and multiple approaches for faculty evaluation (Wolansky, 2001) (Arreola, 2007). Only one study has used a decision analysis model to analyse academic research outputs of individuals (Uzoka, 2008). However, as far as the authors are aware, literature in the area does not provide comprehensive models for the evaluation of academic staff. The availability of literature 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 (Arreola, 2007), is only able to capture a small part of the daily activities of the academic staff and definitely do not cover their performance in research, services and management.

Available evaluation models applied to university units (e.g. faculties) have very often produced different rankings, reflecting key differences on the underlying evaluation models and on methodological options. Most of the evaluation methodologies used in these studies suffer major flaws in both substance and process, and have been criticized by several authors (Turban et al., 2004) (Keeney, See, & Winterfeldt, 2006) (Kao & Pao, 2009). Some of these methodological problems also apply to current evaluation models in use, such as the one based on grid points and used to evaluate academic members in the Spanish university system (Agencia Nacional de Evaluación de la Calidad y Acreditación, 2007). These methods very often incur in the critical mistakes reported in decision analysis literature, including: to treat performance indicators as evaluation criteria, and not distinguishing between the notion of performance and the notion of value; weighting criteria solely on the basis of importance and ignoring the notion of trade-offs underlying additive models (Keeney (1992) calls this the most common critical mistake); and summing up ordinal scores on the criteria giving rise to meaningless overall scores. And, as remarked by (Billaut, Bouyssou, & Vincke, Forthcoming) when reviewing methods used to rank universities, "...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" (pg 1).

3. MULTICRITERIA MODEL

This section explains the choice of modelling approach –a multicriteria decision analysis model- and the key methodological features of the model.

3.1 The overarching framework

The use of multiple criteria methods seems to be an adequate approach to deal with faculty evaluation in complex university systems given that multicriteria methods:

- are grounded on strong theoretical foundations (Keeney & Raiffa, 1976), having the potential to overcome major flaws in both substance and process unfortunately common in previous studies, in particular correcting problems in the selection of the criteria, in the scoring of academic units on these criteria, and in the assessment of the weights for the criteria (Keeney & Raiffa, 1976) (Belton & Stewart, 2001) (Keeney et al., 2006).
- respect the multiple criteria nature of faculty evaluation, given that there might be multiple and conflicting objectives to be considered in faculty evaluation and different viewpoints from individual staff, departments and other stakeholders (Kao & Pao, 2009);
- allow for building comprehensive and transparent models with the same rules applied to all faculty members, deemed as appropriate to overcome problems in evaluation systems which have diffuse purposes, imprecise objectives and criteria and inconsistent administration (Mills & Hyle, 1999) and which are seen as inequitable and might lead to disputes and dissatisfaction among faculty (Elmore, 2008);
- can be designed to respect the socio-technical processes related with evaluation (Phillips, 2007) and be able to combine advanced metrics with transparent peer review (a key feature for faculty evaluation (Moed, 2007)); this feature might contribute for higher acceptability of evaluation systems and enhance the evaluation culture;
- provide problem structuring methods useful to define the multiple and complex strategic objectives to be attained by universities and that should inform faculty evaluation models;
- allow for incorporating in a model the beliefs about academic careers and institutional policies that stakeholders choose to define, as well as quantitative information validated in other studies.

The usefulness of evaluation models based on decision analysis methods, applied in other university contexts, has been identified in (Politis & Siskos, 2004) (Turban et al., 2004) (Keeney et al., 2006) (Uzoka, 2008).

The proposed evaluation approach addresses some methodological aspects relevant to decision analysis literature or not previously addressed within faculty evaluation, namely:

- How to structure the faculty evaluation problem, i.e. how to define an appropriate set of criteria and how to choose descriptors of performance to operationalise the criteria (being aware of issues discussed in (Keeney, 2007))? Which model features allow the application of the model to academics from different scientific areas?
- How to integrate within a multicriteria evaluation model concepts from the literature of human resources and strategic management, such as on management by results? How to build a model that takes into account quantity and quality aspects from academic production?
- Which value functions should be used to convert performance into value, so as to express how important are improvements in performance to the university evaluator? How to construct weights so that they do not reflect directly the importance of criteria, which is the most common mistake in evaluation (Keeney et al., 2006)?
- How to build an evaluation model that respects the autonomy of the faculty profession? E.g., given that within academic autonomy, academic staff might choose to invest more in some academic activities, is there a multicriteria model that is consistent with university objectives and flexible to accommodate different preferences of individuals?

The proposed faculty evaluation model follows a sequence of steps that goes from structuring to evaluation, as defined in (Bana e Costa, Ensslin, Correa, & Vansnick, 1999). The key features and steps for developing the proposed faculty evaluation model are portrayed in Fig. 1.

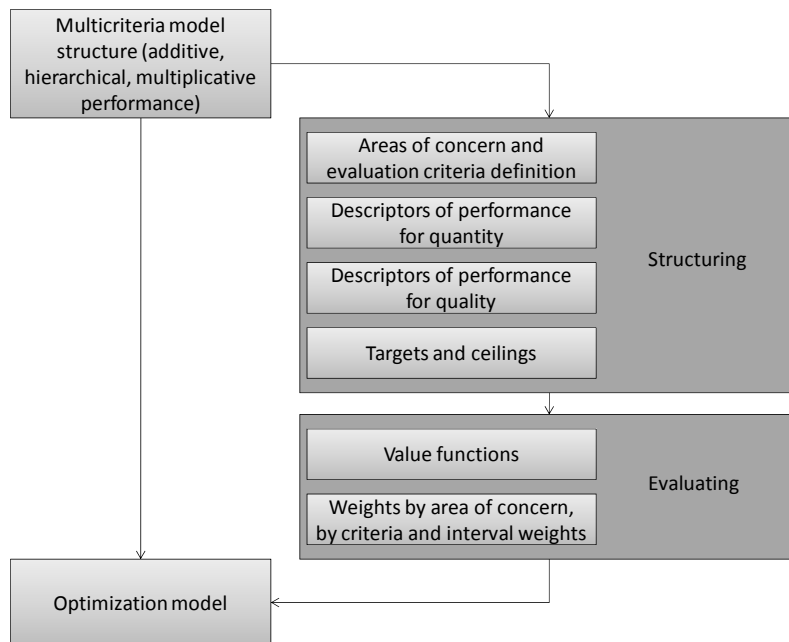


Figure 1: Key components of the faculty evaluation model

At first, the underlying structure of the multicriteria model was defined: it was chosen a two-level hierarchical additive model with the quantitative and the qualitative dimensions being accounted at a bottom non-additive third level (with performance resulting from multiplying quantitative performance by qualitative performance). Secondly, using the structure of the multicriteria model, an optimization model was built. The optimization model chooses the set of weights that maximises the score for an academic member, while respecting the intervals of weights that are consistent with university objectives. Thirdly, in order to apply the multicriteria and the optimization models, the structuring phase involved the choice of areas of evaluation and of evaluation criteria, the choice of descriptors of performance for quantity and for quality, as well as the definition of targets and ceilings. The evaluation phase included the choice of value functions and weights. These model features are discussed in detail in the next sub-section. We assume that the reader is familiar with the theoretical foundations and traditional procedures for the construction of value models in decision analysis (Keeney & Raiffa, 1976) (French, 1986).

3.2 Key features of the model

Additive and nature

A multiattribute additive value model was chosen to convert performance into value. The model has a two-level hierarchical additive structure, and accounts for the quantitative and qualitative dimensions of academic activity at a bottom non-additive third level. The theoretical foundations of additive value models are described in (Keeney & Raiffa, 1976) and (Belton & Stewart, 2001). At the first level of the hierarchy are the key areas for evaluation (that can also be named areas of activity, areas of concern): scientific activity, pedagogical activity, etc. For Portugal, the areas of evaluation for faculty evaluation are normatively defined in the legislation that name the main types of academic activities (Ministério da Ciência Tecnologia e Ensino Superior, 2009). At the second level of the hierarchy are evaluation criteria. These are grouped within the areas for evaluation defined at the first level. The operationalization of evaluation criteria accounts for the quantitative and qualitative dimensions of academic activity –that are defined in a third non-additive level–, which are explained in detail in the quantity vs. quality sub-section.

The mathematical formulation of the two-level hierarchical additive model is as follows –the value of the activity of a faculty member is computed through a weighted sum:

$$V(d) = \sum_j^M \left(\sum_{i_j=1}^{N_j} V_{i_j}^d * w_{i_j} \right) * w_j \quad (1)$$

$$\sum_{i_j=1}^{N_j} w_{i_j} = 1, \forall j \quad (2)$$

$$\sum_{j=1}^M w_j = 1 \quad (3)$$

$$w_{i_j}, w_j \geq 0, \forall i, j \quad (4)$$

In which: d is a faculty member; $V(d)$ is the global value for faculty member d ; $V_{i_j}^d$ represents the partial value obtained by the faculty member on the evaluation criteria i_j that belongs to the j area of evaluation (as shown below, this partial value results from converting performance into value, through the use of a value function); w_{i_j} represents the weight given to evaluation criteria i_j (with weights from evaluation criteria within an area of evaluation summing one, according to Eq. (2)); w_j

represents the weight given to each area of evaluation j (with weights by area of evaluation summing one, as seen in Eq. (3)); and all weights being non-negative (Eq. (4)).

The use of an additive model means that the set of evaluation criteria included in the model needs to be exhaustive, concise, non-redundant and preferentially independent (Keeney, 1992). Given the compensatory nature of an additive model, lower performance in one evaluation criteria might be compensated by a higher performance in another evaluation criteria (nevertheless, there are limits to compensation in academic activity, which has lead to the use of ceilings to performance, later explained).

The V_{ij}^d partial value results from converting performance P_{ij} for the i_j evaluation criteria into value, as defined in Eq. (5). For defining value, the evaluator needs to define proper indicators of performance and to make explicit how much he values improvements in performance.

$$V_{ij}^d = f_{ij}(P_{ij}), \forall i, j \quad (5)$$

Optimization model with interval weighting

Weights are scale constants which capture the trade-offs between the objectives to be attained by universities, e.g. the trade-offs between evaluation criteria. Weights should be defined by the evaluator in accordance to university objectives.

Nevertheless, academic activity integrates a wide set of activities, and different members of faculty staff might prefer to dedicate different attention to different types of activities. Accordingly, we suggest using a flexible weighting system within an optimization model, following a model feature previously used in (Kao & Pao, 2009). Adopting flexible weights means that weights might vary within predefined limits, and for each faculty member, the global value given by the evaluation model will be the maximum value that might be attained given those predefined intervals of weights.

Authors propose the mathematical programming model to maximize the global value for an academic member, defined in Eqs. (6-10), with Eq. (6) being the objective function and Eqs. (7-10) being the constraints defining the intervals of weights¹.

$$Max_{w_{i_j}} \sum_j^M \left(\sum_{i_j=1}^{N_j} V_{i_j}^d * w_{i_j} \right) * w_j \quad (6)$$

Subject to:

$$\overline{w}_{i_j} \leq w_{i_j} \leq \underline{w}_{i_j}, \forall i_j \quad (7)$$

$$\overline{w}_j \leq w_j \leq \underline{w}_j, \forall j \quad (8)$$

$$\sum_{i_j=1}^{N_j} w_{i_j} = 1, \forall j \quad (9)$$

$$\sum_{j=1}^M w_j = 1 \quad (10)$$

With \underline{w}_{i_j} and \overline{w}_{i_j} being the lower and the upper bound for the weights with reference to evaluation criteria; and \underline{w}_j and \overline{w}_j being the lower and upper bound for the weights with reference to the areas of evaluation.

Quantity vs. Quality

When evaluating faculty performance, it is crucial to take into account for both the quantity and the quality of academic production. That is, if the model includes an evaluation criterion of scientific publications, one needs to consider both the quantity and the quality of publications. We needed thus to define the nature of quantity and quality information to be used in the model (in particular, within the descriptors of performance) and the functional form to aggregate quality and quantity in a descriptor of performance. We start by discussing the later².

The chosen multicriteria model had to capture both quantity and quality in each evaluation criteria. Within multicriteria literature, different model structures, using additive and multiplicative features

¹ A general optimization model is presented. Depending on whether weights of intervals are set at the evaluation criteria level, or at the area of evaluation level, the proposed model might be subject to small adjustments.

² A general discussion on the construction of descriptors of performance is available in Bana e Costa, C. A., & Beinart, E. (2005). Model-structuring in public decision-aiding. In O. R. Group (Ed.), *Working paper LSEOR*. London: London School of Economics and Political Science..

are available in (Keeney & Raiffa, 1976; Kirkwood, 1997). A first analysis has shown that an additive relation between quantity and quality would mean that there is a trade-off between the two dimensions, and that potentially one might have zero quantity and high quality, which was shown not to be appropriate (e.g., one can only assess the quality of teaching if there is teaching).

Alternatively, authors found out that, under a set of conditions, a multiplicative relationship between quantity and quality would be appropriate for the context of faculty evaluation: with quality being independent with quantity, and quantity being preferentially dependent with quality. Under this case, the substitution rate between quality and quantity varies with quantity, but not with quality. This model is only valid if and only if the conditions portrayed in Tab. 1 are accepted by the evaluator –in our case, the Board of Directors of IST has accepted those conditions.

By accepting these conditions, when quality is independent with quantity and quantity is preferentially dependent with quality, the evaluator might evaluate quality separately. When one multiplies quality by quantity, it means quantity is adjusted by quality, and the result is converted into quantity units for the same level of quality. Given that performance is the result of multiplying quantity by quality, for each evaluation criteria, performance is defined as:

$$P_{ij} = Q_{ij}^t * Q_{ij}^l, \forall i, j \quad (11)$$

With Q_{ij}^t standing for quantity and Q_{ij}^l standing for quality for an evaluation criteria i_j .

Accordingly, one can read that while the second level of the hierarchical additive model uses information on performance and defines value for each evaluation criteria, quantitative and qualitative performance are defined at a third non-additive level of the faculty evaluation model. Under the selected model one cannot have quality without having a positive quantity. The technical assumptions used in this type of multiplicative structure might be consulted in (Keeney & Raiffa, 1976).

Table 1: Conditions for accepting quantity as preferentially dependent with quality, and quality as independent with quantity (an example applied to a criterion on the scientific production of articles in international journals)

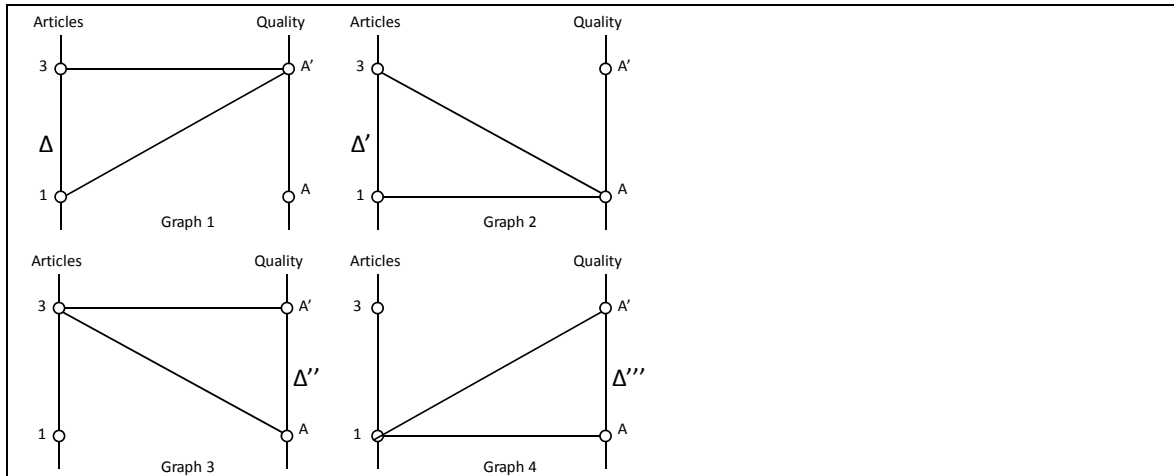


Figure 2: Performance

Question (graph 1 in Fig. 2): How do you define your preference for three articles with high quality (A') in comparison to one article with high quality (A')?

Answer: I consider the difference between the two options to be preferentially strong (Δ).

Question (graph 2 in Fig. 2): How do you define your preference for three articles with low quality (A') in comparison to one article with low quality (A')?

Answer: I consider the difference between the two options as preferentially weak (Δ').

This means that $\Delta > \Delta'$, and that the value attributed to quantity in the production of articles depends on the level of quality.

In addition:

Question (graph 3 in Fig. 2): How do you define your preference for three articles with high quality (A') in comparison to three articles with low quality (A)?

Answer: I consider the difference between the two options as preferentially moderate (Δ'').

Question (graph 4 in Fig. 2): How do you define your preference for one article with high quality (A') in comparison to one article with low quality (A)?

Answer: I consider the difference between the two options as preferentially moderate (Δ''').

This means that $\Delta'' = \Delta'''$, and that the value attributed to quality is independent from quantity; and thus quality can be separately modeled.

When selecting the performance indicators for quality and quantity, one should be aware of the implicit judgments that result from using a multiplicative model, in particular the judgments that result from choosing a descriptor of performance for quality and its corresponding quantitative score. Tab. 2 explains how to make explicit the assumptions which result from multiplying quantity by quality, given the selection of a minimum and of a maximum score for quality performance.

Table 2: Interpreting judgments when one multiplies quantity by quality (an example applied to an evaluation of scientific production, as measured by the number of articles)

Considering an indicator of performance that results from multiplying the number of articles by their quality (for simplicity, we neglect some of the indexes):

$$P = Q^t * Q^l$$

And having:

- two articles a and b with the same quantity $Q^t(a) = Q^t(b)$;
- article a with quality equaling 2 ($Q^l(a) = 2$), with 2 being the maximum level (excellent article);
- article b with quality equaling 0.5 ($Q^l(b) = 0.5$), with 0.5 being minimum level (for an article with very poor quality).

The use of the multiplicative indicator implies that:

$$P(a) = 2Q^t(a)$$

$$P(b) = 0.5Q^t(b)$$

Which means that $P(a) = 4 * P(b)$. This means that excellent article is worth four articles with very poor quality.

Authors propose that quantity and quality descriptors of performance are built with a distinct logic and information basis. Following available literature in the field, whenever possible the descriptor of performance for quantity should make use of objective information and, when compound descriptors of performance need to be built, they should make use of the concept of equivalent unit. Objective information means that it can be computed automatically, does not demand for a specific intervention of the evaluator (Arreola, 2007), and should be informed by related literature (for example, academic studies indicate that citations are a good indicator of performance and citations are correlated with evaluation of content by peers (Meho & Sonnenwald, 2000), meaning that there is a basis for using citations as an indicator of performance within the model).

Building descriptors of performance for objective information should be constructed in line with the concept of *constructed attributes*, as defined by (Keeney, 2007). In many cases, quantifying academic activity demands for accounting for multiple heterogeneous components of activity (such as, for scientific production, summing a scientific article published in a top journal with a scientific article published in a low quality journal). In these cases, authors propose building descriptors of performance based on the equivalent units concept (that has been applied by Keeney et al. (Keeney et al., 2006)). The use of equivalent units demands for building a set of multicriteria sub-models to define the trade-offs across these units (which provide equivalent units coefficients).

Authors propose the use of the MACBETH methodology to build those weights (or trade-off coefficients) (Bana e Costa, De Corte, & Vansnick, 2003; Bana e Costa, De Corte, & Vansnick, Forthcoming). MACBETH stands for Measuring Attractiveness by a Categorical Based Evaluation Technique and is a qualitative approach for scoring options and to weight criteria that requires for qualitative pairwise comparison judgements of difference in attractiveness (value) (Bana e Costa & Chagas, 2004). The MACBETH methodology has been extensively applied in various contexts, including to the evaluation context (Bana e Costa & Vansnick, 1997) (Bana e Costa, Correa, De Corte, & Vansnick, 2002) (Clivillé, Berrah, & Mauris, 2007) (Bana e Costa, Lourenço, Chagas, & Bana e Costa, 2008), to the human resources area (Bana e Costa & Chagas, 2004) and to the supplier evaluation and management context (Oliveira & Lourenço, 2002). The mathematical foundations of MACBETH are described in (Bana e Costa & Vansnick, 1997) (Bana e Costa et al., 2003) (Bana e Costa et al., Forthcoming), and its application is visually supported by the M-MACBETH software (Bana e Costa et al., 2003).

With regard to quality, authors propose that descriptors of performance for quality make use of subjective information, while respecting the principle of “controlled subjectivity” proposed by Arreola (2007). The use of subjective information means that assessment is done by the evaluator (being a form of peer review) and should capture behavioral, human, strategic and other aspects relevant to the evaluation of human resources. According to the principle of “controlled subjectivity”, the form of computation of quality performance should be clear and should make use of scales and methodological choices indicated by literature (National Academy of Engineering, 2009). Following guidelines from literature (Arreola, 2007) (National Academy of Engineering, 2009), all the descriptors of performance for quality should use the same performance scale, i.e., a single scale with the same number of points should be used for all the evaluation criteria, contributing to consistency across

measurements. Authors propose the following procedure to reach a performance scale for each evaluation criteria:

1. Definition of a list of pros and cons associated with the production of a quantitative element;
2. Converting the list of pros and cons into a pre-defined performance scale and a score that operates as a correction to quantity.

When building performance indicators for quantity and quality, ratio scales were used. Ratio scales have good properties (a detailed analysis on their properties is available in (Stevens, 1946)), allow for using the zero as an absolute reference, and are compatible with the types of units commonly used in quantitative performance in faculty evaluation. For example, the use of ratio scales allows for stating that the publication of an international book is worth five times the publication of an article in an international journal (note that one cannot interpret that international books are five times more important or five times bigger than articles in international journals).

Using performance targets and ceilings

Within the faculty evaluation context, authors found useful to define references in performance that are meaningful and useful for the management of human resources. The option was to use targets, a concept that is commonly used in strategy and human resources management literature (Drucker, 1993). Thus, within the multicriteria model, for each indicator of performance, P_{ij} , the evaluator must define a performance target that is indicative of good performance. Given the use of ratio scales and of performance targets, authors have selected the following reference points for building weights: the zero and the target, having respectively values of 0 and 100 (Eq.s (12) and (13)). These reference points are meaningful to the evaluator, making easier the weighting task.

$$P_{ij} = 0 \Rightarrow V_{ij}^d = 0 \quad (12)$$

$$P_{ij} = target_{ij} \Rightarrow V_{ij}^d = 100 \quad (13)$$

Also, given that in the context of faculty evaluation there might be limits to the compensatory nature of a multicriteria model, e.g., the evaluator might value additional activity in one evaluation criteria up to a specific limit. Limiting value for one academic activity is important in terms of incentives: if there are no limits, one academic member might only carry out that activity, disregarding any other

activities. This has led to the utilization of ceilings within the model. The use of a ceiling means that, for each evaluation criteria, there is a point after which further activity by one faculty member does not contribute for an increase in value. Authors propose that the use of a ceiling is modeled through the choice of a value function with special characteristics, like the one proposed in the next subsection.

Building value functions and weights

Value functions convert performance into value and should account for the preferences of the evaluator and for the incentives promoted by the evaluation model. Authors propose, for all the evaluation criteria, the use of an S-shaped value function with the format presented in Fig. 3, which is both compatible with the use of ceilings, which respects the delta property and which signals the importance of targets. An S-shaped value function tends to reward performance close to the target. In case of performance below the target, marginal increases in performance are gradually more valued by the evaluator, and in case of performance above the target, marginal increases in performance are valued at a decreasing marginal rate.

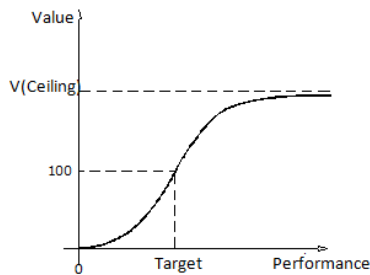


Figure 3: S-shaped value function

A specification for the S-shaped value function is shown in Eq. (14). For each evaluation criteria, a different value function needs to be built (for simplicity, we omit the index i_j in the equations). Two branches are defined by exponential functions, whereas the exponential function respects the delta property, also named the constant trade-off attitude condition (Kirkwood, 1997). The constant trade-off attitude condition holds when, for an additive value function, for an attribute X_i , if it is true that x_i^m is the midvalue of $[x_i', x_i'']$, then $x_i^m + \delta$ is the midvalue of $[x_i' + \delta, x_i'' + \delta]$, for any δ (the proof and the rationale for the constant trade-off condition are presented in (Kirkwood, 1997)). Authors consider this is a desirable property in the context of faculty evaluation.

$$V_{ij} = V_{ij}(P) = \begin{cases} \frac{1 - \exp[-P/\rho_1]}{1 - \exp[-P_{target}/\rho_1]}, & 0 \leq P \leq P_{target} \\ \frac{1 - \exp\left[-\frac{(P - P_{target})}{\rho_2}\right]}{1 - \exp\left[-\frac{(P_{ceiling} - P_{target})}{\rho_2}\right]}, & P_{target} \leq P < P_{ceiling} \\ V_{ceiling}, & P \geq P_{ceiling} \end{cases} \quad (14)$$

With: P_{target} being the performance target; $P_{ceiling}$ being the ceiling defined in terms of performance; $V_{ceiling}$ being the ceiling defined in terms of value; and ρ_1 and ρ_2 parameters that determine the shape of the exponential branches and need to be selected by the evaluator. ρ_1 is negative, and the lower its value, the closer the value function will be to a linear function; ρ_2 is positive, and the higher its value, the closer the value function will be to a linear function.

The use of inappropriate procedures to build weights in a multicriteria model is the most common critical mistake in decision analysis (Keeney, 1992), and a common problem in evaluation studies (Keeney et al., 2006). Weights capture not only the importance of the evaluation criteria, but also how important are the differences between the defined reference points. The choice of weights should respect the methodological orientations defined in the decision analysis literature (von Winterfeldt 1998) (Bana e Costa, Corrêa et al. 2002) (Keeney 2002). Building weights for a hierarchical model demands awareness of several issues related with the construction of weights generally, and with the construction of weights in hierarchical models in particular (Weber, Eisenfurh, & Von Winterfeldt, 1988) (Borcherding & von Winterfeldt, 1988) (Weber & Borcherding, 1993) (Lourenço, 2002).

The proposed model demands for elicitation of weights at two hierarchical levels (by area of evaluation and by evaluation criteria) and of intervals of weights. The following methodology is proposed for building weights:

1. Use of the MACBETH methodology to build a first set of point weights;
2. Use of swings to compute rounded weights and to calculate the maximum and the minimum 'acceptable' swings on weights across areas of evaluation and/or evaluation criteria;

- Use of those swings to build intervals of weights as defined in Eq.s (14) and (15) (maximum and minimum limits, respectively) that might afterwards be adjusted by the evaluator.

$$\overline{w}_{i_j} = \frac{s^{max_{i_j}}}{s^{max_{i_j}} + \sum_{y \neq i_j} s_y^{min}} \quad (15)$$

$$\underline{w}_x = \frac{w^{min_{i_j}}}{s^{min_{i_j}} + \sum_{y \neq i_j} s_y^{max}} \quad (16)$$

With: $s^{min_{i_j}}$ being the minimum swing for i_j , $s^{max_{i_j}}$ the maximum swing for i_j and swings defined with reference to academic staff profiles (swings are non-normalized weights built by comparison to a reference –examples are presented in the IST model section).

Application by scientific area and socio-technical processes

The multicriteria model is generally built and is flexible for application in different scientific areas. While some of the values to be used in the model should be fixed, others should be defined for each scientific area. We consider that the weights should be fixed for the entire faculty, while some targets, ceilings and other specific information of the model might be defined for each scientific area (other information might include the choice of which are the ‘top journals’ in the area). Nevertheless, one should note that the use of different targets and ceilings requires specific analysis to ensure that the model is well calibrated.

This model was developed with the help of two decision analysts that assumed the role of facilitators/analysts – CBC and MO –, and with a group of IST stakeholders, with PM and AS assuming the role of technical advisors and representatives of key stakeholders. The development of the model at IST has also benefited from feedback from the school, given that different versions of the model were circulated and made available for public discussion.

The multicriteria faculty evaluation model not only produces an overall value for a faculty member, but also generates a wide set of key information for university management. Namely, it allows for defining the multicriteria value profile of each faculty member, as shown in Fig. 4. Furthermore, using the information generated by the multicriteria faculty evaluation model, other profiles might be analyzed, e.g., the multicriteria profile of each faculty member by area of evaluation and the typical multicriteria profile by scientific area.

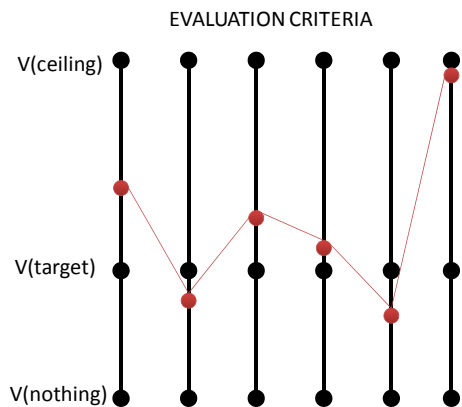


Figure 4: Multicriteria value profile of a faculty member (in red)

There are many issues related with the social process required for the application of the model. It is outside the scope of this study to discuss these issues, although we raise some related issues in the conclusions section.

4. IST MODEL

IST has adopted the proposed faculty evaluation model to evaluate its academic staff in the 2010-2012 period. Guiding rules defining the generic model to be used in evaluation have been discussed and approved within the school and in different boards (Universidade Técnica de Lisboa, 2010). This section presents key features of the model adopted by IST, as well as some examples of decision analysis tools resulting from the application of the faculty evaluation model.

4.1 Structure

The selected hierarchical model is depicted in Fig. 5. At the first level of the hierarchy there are four areas of evaluation –teaching, research, knowledge transfer and university management–, and most areas of evaluation include several evaluation criteria (evaluation criteria are shown in red in Fig. 5). These areas of evaluation and evaluation criteria entail the strategic objectives to be attained by IST. Each evaluation criteria was made operational through the construction of descriptors of performance for quantity and quality (third hierarchical level in Fig. 5).

For each evaluation criteria IST has chosen a different descriptor of performance for quantity, while a unique descriptor of performance was adopted for quality, being this descriptor applied to each evaluation criteria. In what follows authors explain how the descriptor of performance in international publications was built using the concept of equivalent units and a multicriteria sub-model to compute the trade-off ratios required to sum equivalent units.

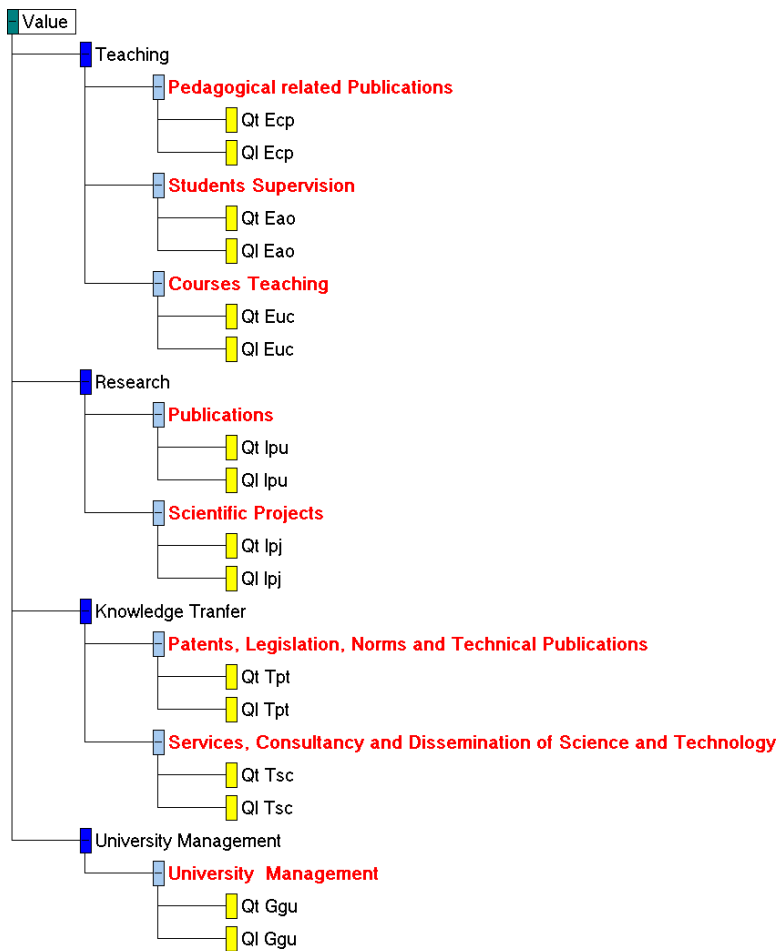


Figure 5: Hierarchical value tree

The descriptor of performance for quantity was defined as:

$$Q_{Publications_{Research}}^t = \sum_{k=1}^K \frac{1}{Z_k} * \left(T_k * \theta_k + \frac{1}{\rho} R_k \right) \quad (16)$$

With: k being a scientific and international publication in the evaluation period ($k = 1, \dots, K$); T_k being the factor that converts publications into a common equivalence unit; θ_k being a coefficient adjusting for the age of the publication; ρ standing for the reference number of citations in the scientific area (for a specific period of time); R_k being the number of citations of the publication, excluding self-citations; and Z_k being a correction factor for the number of authors.

We focus here on how the MACBETH methodology was used to build equivalent units for T_k . Fig. 6 presents the matrix filled with comparisons from different types of publications. One might read in the matrix that the difference in attractiveness (value) between publishing an international book and publishing an article in a Type A journal is extreme, and the difference between publishing an article in a type C journal and a chapter in a book is very strong. Following this set of qualitative judgments, the M-MACBETH software assisted in proposing a quantitative value function that might be adjusted and should be validated by the evaluator. After discussion, IST decided to approve the equivalent units that are listed in Tab. 3.

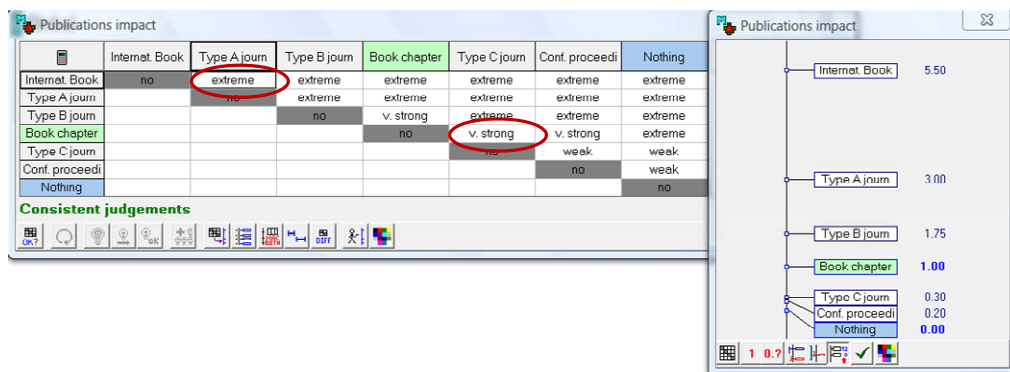


Figure 6: MACBETH judgments for different types of scientific international publications

Table 3: Equivalent units for different types of scientific and international 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

IST has adopted descriptors of performance for quality on a common scale that is to be consistently applied to all evaluation criteria. For each staff member and for each evaluation criteria, the evaluator needs to justify the qualitative performance using the following procedure:

- List pros ('strong points') and cons ('weak points') for the activity of the staff member;
- Classify the pros and cons as 'determinant' or 'non-determinant', and justify that classification;
- Use Tab. 4 to define the level of performance and the corresponding score.

Table 4: Descriptors of performance for the qualitative indicators

Levels of performance	Description	Q^l
Highly positive	<i>There is at least one 'strong' determinant point and no 'weak' determinant points</i>	1.5
Positive	<i>There are no 'strong' nor 'weak' determinant points, and 'strong' points more than compensate 'weak' points</i>	1.25
Neutral	<i>There was no identification of neither 'strong' nor 'weak' determinant points, or the 'strong' points are balanced with 'weak' points</i>	1.0
Negative	<i>There are no 'strong' nor 'weak' determinant points, and 'weak' points more than compensate 'strong' points</i>	0.75
Highly negative	<i>There is at least one 'weak' determinant point and no 'strong' determinant points</i>	0.5

With regard to targets, IST decided that some targets should be defined by scientific area, and issued examples of targets, such as the ones listed in Tab. 5. Targets should be seen as levels of performance to be attained by academic staff, and the reasoning why some targets might vary across scientific areas is the following: with regard to scientific production, the average and the expected number of publications by individual structurally differs, being thus fair to consider a different target across scientific areas.

Table 5: Activity targets for a three year period

Evaluation criteria	i_j	Target	Examples of targets (to be interpreted with $Q^l = 1$)
Pedagogical related publications	Ecp	1.5	1 book chapter and 1 pedagogical text
Students supervision	Eao	6	2 supervision of MSc thesis per year
Courses teaching	Euc	9	9h of teaching courses per week with normal evaluation by students
Publications	lpu	4.5	2 articles in type B journals and 1 chapter in an international book
Scientific projects	lpj	1	Responsibility for 1 national R&D project
Patents, legislation,...	Tpt	9	1 international patent
Services, consultancy,...	Tsc	6	12 service or consultancy actions with reduced funding
University management	Ggu	4.5	Being the dean of studies of an MSc course

The ceilings selected by IST are presented in Tab. 6. When defining the ceilings, attention should be paid to the performance of a faculty member that attains the ceiling in one criterion and has zero performance in all the other criteria.

Table 6: Ceilings defined in terms of value

Evaluation criteria	Pedagogical related publications	Students supervision	Courses teaching	Publications	Scientific projects	Patents, legislation,...	Services, consultancy,...	University management
i_j	Ecp	Eao	Euc	lpu	lpj	Tpt	Tsc	Ggu
$V(\text{ceiling})$	500	300	300	600	500	300	300	500

4.2 Value functions and weights

For simplicity, IST has chosen to adopt piecewise linear functions for the 2010-2012 period. An example of a piecewise linear value function, alternative to the S-shaped value function is shown in Fig. 7 (for the publications evaluation criteria). Calibration of piecewise linear value functions demands for information on the target and on the ceiling (in terms of performance or in terms of value), while a calibration of a S-shaped function would demand for further information from the evaluator.

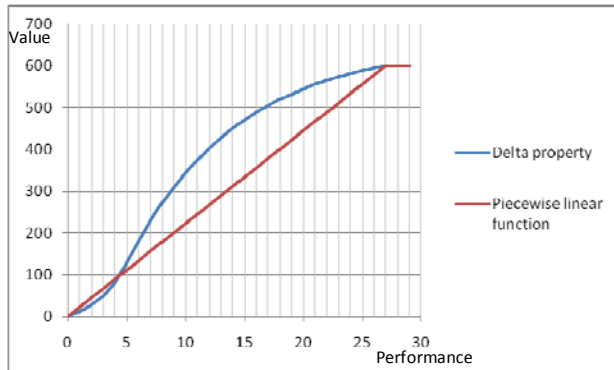


Figure 7: Value function chosen by IST vs. an alternative S-shaped value function, for the publications evaluation criteria

Fig. 8 illustrates how a multicriteria profile for an IST faculty member can be presented, given the targets and ceilings chosen by IST.

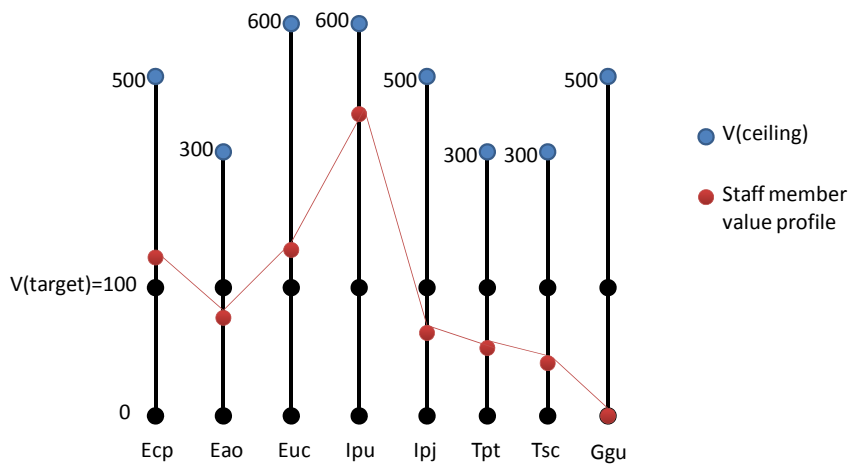


Figure 8: Multicriteria value profile for a staff member

With regard to weights, a first approximation of weights was computed using the MACBETH methodology. This has meant making pairwise qualitative comparisons of fictitious profiles of academic staff, as shown in Fig. 9. For example, the following judgments were required: the difference between a staff member that has achieved the target for publications and has not carried out any other activity (Ipu) and a staff member without activity (All being zero) is very strong; and the difference between a staff member that has achieved the target for publications and has not carried out any other activity (Ipu) and a staff member that has achieved the target in teaching and has not

carried out any other activity (Ecu) is strong. The M-MACBETH software was used to analyze the set of quantitative weights that were compatible with the qualitative judgments (values also presented in Fig. 9).

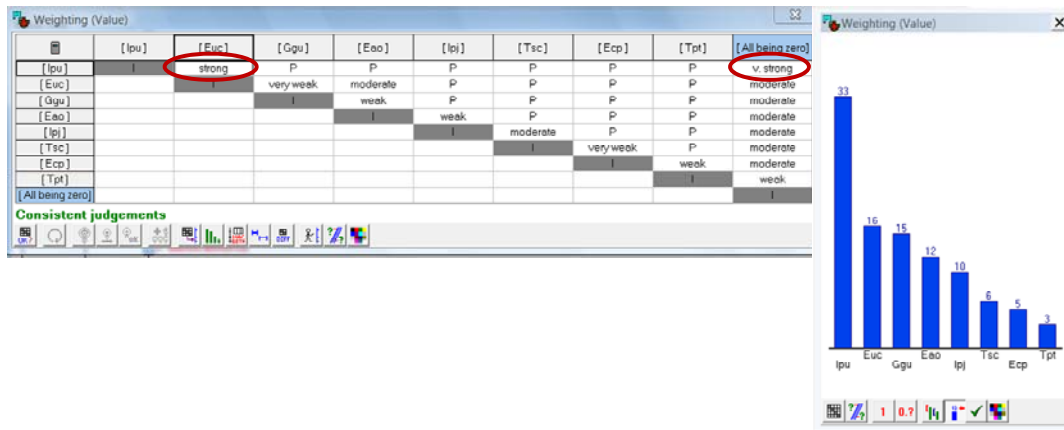


Figure 9: First set of weights obtained using the MACBETH methodology (names of evaluation criteria can be found in Fig. 5)

In a second stage, the weights were analyzed within the hierarchical structure, and were re-computed through the use of swings used to get rounded values, as shown in Tab. 7. The basic unit used to compute swings was 1 (which applied to the Ecp and Tpt evaluation criteria). Swing multipliers of 1 were defined for all the other evaluation criteria, and the sum of swings for criteria within each area of evaluation gave the swings for the areas of evaluation –e.g., the swing 6 for the pedagogical area of evaluation resulted from summing the swings for the respective evaluation criteria, $3(Ecp)+2(Ecu)+1(Eao)$.

Table 7: Weights by area of evaluation and by evaluation criteria (names of evaluation criteria can be found in Fig. 5)

Area of evaluation	Pedagogical			Research		University management	Knowledge transfer	
Evaluation criteria	Ecu	Eao	Ecp	Ipu	Ipi	Ggu	Tsc	Tpt
Scores (and respective w_j)	6 (30%)			8 (40%)		3 (15%)	3 (15%)	
Scores (and w_{ij} normalised to sum 100%)	3 (15%)	2 (10%)	1 (5%)	6 (30%)	2 (10%)	3 (15%)	2 (10%)	1 (5%)

In the last stage, building intervals of weights required the choice of an additional set of swings. IST has chosen to define intervals of weights at the area of evaluation level (and not at the evaluation criteria level). The second column in Tab. 8 presents the (point) swings used in Tab. 7, after these being normalized so that the highest weight is 100 (the reference is built against the research area of evaluation). Tab. 8 presents the minimum and maximum swings that were computed using fictitious profiles similar to the ones defined above using the MACBETH method. One might read values in Tab. 8 as follows: when a shift from a staff member that has no activity to a staff member that reaches the targets in all the evaluation criteria within the research area of evaluation is valued 100, a shift from a staff member that has no activity to a staff member that reaches the targets in the all the evaluation criteria within the pedagogical area of evaluation cannot be valued as less than 50% or as more than 90% of that value (the minimum and the maximum swings). Applying Eq.s (15-16) to the swings, one obtains the intervals of weights presented in the last two columns of Tab. 8.

Table 8: Swings generated so as to define the intervals of weights

	Swing computed from Tab. 7	Min. swing	Max. swing	Min. weight	Max. weight
Pedagogical	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%

Those intervals of weights were further adjusted, and the final weights selected by IST to evaluate assistant professors are listed in Tab. 9. Given that Portuguese legislation attributes different management responsibilities to academic staff along the academic hierarchy, the minimum weight for university management for associate and full professors was higher than 0% (2,5% and 5%, respectively).

Table 9: Set of weights adopted by IST

Area of evaluation		Pedagogical			Research		Knowledge transfer		University management
Weight by area of evaluation	w_j	20%-40%			40%-60%		0%-20%		0%-20%
Evaluation criteria	i_j	Ecp	Eao	Euc	lpu	lpj	Tpt	Tsc	Ggu
Weight by evaluation criteria	w_{i_j}	1/6	2/6	3/6	6/8	2/8	1/3	2/3	1
Weight by evaluation criteria (normalized)	—	6.7%	13.3%	20%	45%	15%	6.7%	13.3%	20%

4.3 Model requisiteness

IST has adopted the optimization model earlier presented and that makes use of the information of the multicriteria just presented.

The values in use in the multicriteria model at IST were subject to simulations and tests, so as to ensure that the model was well calibrated and adequately distinguishing between different types of staff members. These simulations were critical to ensure the requisiteness of the model, i.e., that the model is enough in content and form so that it could be used within an evaluation system (Phillips, 1984).

The use of the model for evaluation at IST has required additionally the definition of thresholds to classify staff members within evaluation categories (being this classification required by law). The definition of these thresholds was obtained using a set of academic profiles and with the respective computation of the global value. Examples of the initial profiles used and respective value computations are presented in Tab. 10. After testing multiple profiles, categories and thresholds at IST were set according to Eq. (17).

$$Category\ assignment^d = \begin{cases} 'Inadequate', 0 \leq V(d) < 20 \\ 'Sufficient', 20 \leq V(d) < 40 \\ 'Relevant', 40 \leq V(d) < 80 \\ 'Excellent', V(d) \geq 80 \end{cases} \quad (17)$$

Table 10: Selected profiles used to compute category thresholds

Profile 1	A staff member that in the evaluation period has taught 9 hours per week/semester with students' evaluation of 5, with a quality performance of 1 and not having carried out any other activity, will have $V(d) = 0.2 * V_{Euc}(9) = 0.2 * 100 = 20$ Category: 'Sufficient'.
Profile 2	A staff member that in the evaluation period has taught 6 hours per week/semester, with a students' evaluation of 5 and a quality performance of 1, with 3 articles published in type C journals, with 1 co-author, without citations and with a quality performance of 1, and not having carried out any other academic activity, will have $V(d) = 0.2 * V_{Euc}(6) + 0.45 * V_{Ipu}(3) = 0.2 * 66.7 + 0.45 * 66.7 = 43$ Category: 'Relevant'.
Profile 3	A staff member that in the evaluation period has been the head of the department, with a quality performance of 1, and not having carried out any other academic activity will have $V(d) = 0.2 * V_{Ggu}(9) = 0.2 * 200 = 40$ Category: 'Relevant'.
Profile 4	A staff member that in the evaluation period has occupied management responsibilities of 13.5 hours per semester with a quality performance of 1, and not having carried out any other academic activity will have $V(d) = 0.2 * V_{Ggu}(13.5) = 0.2 * 300 = 60$

	Category: 'Relevant'.
Profile 5	A staff member that in the evaluation period has taught 6 hours per week/semester, with a students' evaluation of 5 and a quality performance of 1, with 3 articles published in type B journals, with 1 co-author, without citations and with a quality performance of 1, with supervision of three MSc thesis with a quality performance of 1 and not having carried out any other academic activity, will have $V(d) = 0.2 * V_{Ecu}(6) + 0.133 * V_{Eao}(3) + 0.45 * V_{Ipu}(6) = 0.2 * 66.7 + 0.133 * 50 + 0.45 * 133 = 80$ Category: 'Excellent'.

5. CONCLUSIONS

The adoption of a faculty evaluation model was viewed at IST as an opportunity for improving current practices in place. IST has already formally decided to adopt the proposed faculty evaluation model, expecting to establish an evaluation system that fairly treats staff members from different scientific areas and that motivates staff to re-align their activities in accordance to the school strategic goals.

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

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 evaluation of teaching be done only by students, or peers should also review the content of teaching? If peer review of the context of teaching is required, which methods for evaluation are available and have been validated? Which is the best indicator for the impact of publications on the community? Moreover, the model was developed without detailed information on the performance of IST academic staff in many indicators. The use of high-quality data and of scientific metrics might contribute to build a sounder model and to a higher level of acceptance. 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 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 in the 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 thresholds be linked with targets and with ceilings? Last but not least, the proposed model should be tested within schools other than engineering, to investigate whether it respects the evaluation context in other scientific areas. In fact, one should be aware that disciplines might differ in their preferred approach to knowledge transmission.

REFERENCES

- Adler, N. J., & Harzing, A. W. (2009). When knowledge wins: Transcending the sense and nonsense of academic rankings. *Academy of Management Learning and Education*, 8(1), 72-95.
- Agencia Nacional de Evaluación de la Calidad y Acreditación (2007). Programa de Evaluación de Profesorado para la Contratación: Principios y Orientaciones para la Aplicación de los Criterios de Evaluación. Agencia Nacional de Evaluación de la Calidad y Acreditación.
- Arreola, R. A. (2007). *Developing a Comprehensive Faculty Evaluation System: A Guide to Designing, Building, and Operating Large-Scale Faculty Evaluation Systems (3rd ed.)*: Anker Publishing Company.
- Bana e Costa, C., Lourenço, J. C., Chagas, M. P., & Bana e Costa, J. C. (2008). Development of reusable bid evaluation models for the Portuguese Electric Transmission Company. *Decision Analysis*, 5(1), 22-42.
- Bana e Costa, C. A., & Beinart, E. (2005). Model-structuring in public decision-aiding. In O. R. Group (Ed.), *Working paper LSEOR*. London: London School of Economics and Political Science.
- Bana e Costa, C. A., & Chagas, M. P. (2004). A career choice problem: an example of how to use MACBETH to build a quantitative value model based on qualitative value judgments. *European Journal of Operational Research*, 153(2), 323-331.
- Bana e Costa, C. A., Correa, E., De Corte, J. M., & Vansnick, J. C. (2002). Facilitating bid evaluation in public call for tenders: a socio-technical approach. *Omega*, 30, 227-242.
- Bana e Costa, C. A., De Corte, J.-M., & Vansnick, J.-C. (2003). MACBETH. In O. R. Department (Ed.), *Working Paper LSEOR 03.56*. London: London School of Economics of Political Science.
- Bana e Costa, C. A., De Corte, J.-M., & Vansnick, J.-C. (Forthcoming). MACBETH (Measuring Attractiveness by a Categorical Based Evaluation Technique). In J. J. Cochran (Ed.), *Wiley Encyclopedia of Operations Research and Management Science*: Wiley.
- Bana e Costa, C. A., Ensslin, L., Correa, E. C., & Vansnick, J.-C. (1999). Decision Support Systems in action: Integrated application in a multicriteria decision aid process. *European Journal of Operational Research*, 113, 315-335.
- Bana e Costa, C. A., & Vansnick, J. C. (1997). Applications of the MACBETH approach in the framework of an additive aggregation model. *Journal of Multicriteria Decision Analysis*, 6(2), 107-114.
- Belton, V., & Stewart, T. J. (2001). *Multiple Criteria Decision Analysis: An Integrated Approach*: Springer.
- Billaut, J.-C., Bouyssou, D., & Vincke, P. (Forthcoming). Should you believe in the Shanghai ranking? An MCDM view. *Scientometrics*.

- Borcherding, K., & von Winterfeldt, D. (1988). The effect of varying value trees on multiattribute evaluations. *Acta Psychologica*, 68, 153–170.
- Campion, W. J., Mason, D. V., & Erdman, H. (2000). How Faculty Evaluations Are Used In Texas Community Colleges. *Community College Journal of Research and Practice*, 24, 169–179.
- Clivillé, V., Berrah, L., & Mauris, G. (2007). Quantitative expression and aggregation of performance measurements based on the MACBETH multi-criteria method. *International Journal of Production Economics*, 105 171–189.
- Coccia, M. (2008). Measuring scientific performance of public research units for strategic change. *Journal of Informetrics*, 183-194.
- Desselle, S. P., Mattei, T. J., & Vanderveen, R. P. (2004). Identifying and Weighting Teaching and Scholarship Activities Among Faculty Members. *American Journal of Pharmaceutical Education*, 68(4).
- Donovan, C. (2007). The qualitative future of research evaluation. *Science and Public Policy*, 34(8), 555-563.
- Drucker, P. (1993). *Management by results*: Collins.
- Elmore, H. W. (2008). Toward Objectivity in Faculty Evaluation. *Academe*, 94(3), 38-40.
- Etzkowitz, H. (2003). Research groups as 'quasi-firms': the invention of the entrepreneurial university. *Research Policy*, 32(1), 109-121.
- French, S. (1986). *Decision Theory: An Introduction to the Mathematics of Rationality*. Chichester: Ellis Horwood.
- Giménez, V. M., & Martínez, J. L. (2006). Cost efficiency in the university: A departmental evaluation model. *Economics of Education Review*, 25, 543–553.
- Grant, J., & Fogarty, T. (1998). Faculty evaluation as a social dilemma: a game theoretic approach. *Accounting Education* 7(3), 225-248.
- Huber, M. T. (2002). Faculty Evaluation and the Development of Academic Careers. *New Directions for Institutional Research*, 114, 73-83.
- Kao, C., & Pao, H.-L. (2009). An evaluation of research performance in management of 168 Taiwan universities. *Scientometrics*, 78(2), 261-277.
- Keeney, R. L. (1992). *Value-focused thinking: A Path to Creative Decisionmaking*: Harvard University Press.
- Keeney, R. L. (2007). Developing Objectives and Attributes. In W. Edwards, R. F. Miles Jr., & D. von Winterfeldt (Eds.), *Advances in Decision Analysis: From Foundations to Applications*. Cambridge: Cambridge University Press, 104-128.

Keeney, R. L., & Raiffa, H. (1976). *Decisions With Multiple Objectives: Preferences and Value Tradeoffs*. New York: Wiley.

Keeney, R. L., See, K. E., & Winterfeldt, D. v. (2006). Evaluating Academic Programs: With Applications to U.S. Graduate Decision Science Programs. *Operations Research*, 54(5), 813–828.

Kirkwood, C. W. (1997). *Strategic decision making: Multiobjective decision analysis with spreadsheets*. Belmont, California: Duxbury Press.

Lane, J. (2010). Let's make science metrics more scientific. *Nature*, 464, 488-489.

Lourenço, J. C. (2002). Modelo aditivo hierárquico: exemplos de métodos de ponderação e problemas associados. *Working Paper 13-2002*: CEG-IST.

Meho, L. I., & Sonnenwald, D. H. (2000). Citation Ranking Versus Peer Evaluation of Senior Faculty Research Performance: A Case Study of Kurdish Scholarship. *Journal of the American Society for Information Science*, 51(2), 123-138.

Mills, M., & Hyle, A. E. (1999). Faculty evaluation: A prickly pair. *Higher Education*, 38, 351-371.

Ministério da Ciência Tecnologia e Ensino Superior (2009). Decreto-Lei n.º 205/2009. *Diário da República*, 1.ª série(168).

Moed, H. F. (2007). The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science and Public Policy*, 34(8), 555-563.

Murias, P., Miguel, J. C. d., & Rodríguez, D. (2008). A Composite Indicator for University Quality Assesment: The Case of Spanish Higher Education System. *Social Indicators Research*, 89, 29–146.

National Academy of Engineering (2009). *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, 52 pages.

Oliveira, R. C., & Lourenço, J. C. (2002). A multicriteria model for assigning new orders to service suppliers. *European Journal of Operational Research*, 139(2), 390-299.

Phillips, L. (2007). Decision conferencing. In W. Edwards, R. F. Miles, & D. von Winterfeldt (Eds.), *Advances in Decision Analysis: From Foundations to Applications*: Cambridge University Press, 375-399.

Phillips, L. D. (1984). A theory of requisite decision models. *Acta Psychologica*, 56, 29-48.

Politis, Y., & Siskos, Y. (2004). Multicriteria methodology for the evaluation of a Greek engineering department. *European Journal of Operational Research* 156, 223-240.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677-684.

Turban, E., Zhou, D., & Ma, J. (2004). A group decision support approach to evaluating journals. *Information & Management*, 42, 31–44.

Universidade Técnica de Lisboa (2010). Regulamento de Avaliação de Desempenho dos Docentes do Instituto Superior Técnico. *Diário da República*, 51(12128-12137).

Uzoka, F.-M. E. (2008). A fuzzy-enhanced multicriteria decision analysis model for evaluating university Academics' research output. *Information Knowledge Systems Management* 7, 273–299.

Von Winterfeldt, D., & Edwards, W. (1986). *Decision Analysis and Behavioral Research*: Cambridge University Press.

Weber, M., & Borcherding, K. (1993). Behavioral influences on weight judgments in multiattribute decision making. *European Journal of Operational Research*, 67, 1-12.

Weber, M., Eisenfurh, F., & Von Winterfeldt, D. (1988). The effects of splitting attributes on weights in multiattribute utility management. *Management Science*, 34(4), 431-445.

Wolansky, W. D. (2001). A Multiple Approach to Faculty Evaluation. *Education*, 97(1), 81-96.