Evaluating Higher Education Teaching Performance using Combined Analytic Hierarchy Process and Data Envelopment Analysis

Emmanuel Thanassoulis[a], Prasanta Kumar Dey[a,1], Konstantinos Petridis[a], Ioannis Goniadis[b], Andreas C. Georgiou[b]

[a]Operations & Information Management Group, Aston Business School, Aston University
Aston Triangle, Birmingham, UK

[b]Department of Business Administration, University of Macedonia, 156 Egnatia Street, 54636, Thessaloniki, Greece

Disclaimer: This is an earlier version of this paper, forthcoming (2017) in the Journal of the Operational Research Society and so it may differ in certain places from the published version.

[1] Corresponding author: Prof. Prasanta Dey, OIM Group, Aston Business School, Aston University, Birmingham B4 7ET, UK, p.k.dey@aston.ac.uk

Evaluating Higher Education Teaching Performance using Combined Analytic Hierarchy Process and Data Envelopment Analysis

Abstract

Evaluating higher education teaching performance is complex as it involves consideration of both objective and subjective criteria. The Student Evaluation of Teaching (SET) is used to improve higher education quality. However, the traditional approaches to considering students' responses to SET questionnaires for improving teaching quality have several shortcomings. This study proposes an integrated approach to higher education teaching evaluation that combines the Analytical Hierarchy Process (AHP) and Data Envelopment Analysis (DEA). The AHP allows consideration of the varying importance of each criterion of teaching performance while DEA enables the comparison of tutors on teaching as perceived by students with a view to identifying the scope for improvement by each tutor. The proposed teaching evaluation method is illustrated using data from a higher education institution in Greece.

Keywords: Student Evaluation of Teaching; DEA; AHP; Higher Education; SET Questionnaires

1. Introduction

The notion of evaluation goes many years back (DuBois and Garson, 1970; Lincoln and Guba, 1981;Theofilidis, 1989; Tsimboukis, 1979). However, as a scientific field it has started to develop in the mid 60's in the United States and in Great Britain (Worthen and Sanders, 1987). The idea of evaluation is broader and concerns a variety of areas. A general definition of evaluation is the following: Evaluation is the recognition, clarification and implementation of basic criteria in order to define the value of an object based on those criteria (Fitzpatrick et al, 2004).

With globalization of higher education, universities have become part of the services industry and in order to remain competitive university management has become increasingly concerned with students' satisfaction. Students' satisfaction with teaching and learning is considered as one of the major criteria for today's university rankings (Douglas et al, 2006). Therefore, Student Evaluation of Teaching (SET) questionnaires are increasingly used by higher education institutions to evaluate and improve teaching performance of individual teachers. The SET in conjunction with experts' opinion may also be used to formulate strategies for enhancing students' experience. Typically, SET forms serve as formative and summative evaluation that are used by higher education management to improve teaching and learning effectiveness, provide information relevant to promoting staff members and enhancing students' overall experience in specific courses and subjects (Gray and Bergmann, 2003). The purposes of evaluating teaching performance are to develop each teacher's professionalism, to encourage self-improvement, and to maintain achievements (Chen et al, 2015).

Evaluation of teaching performance is challenging as the criteria for evaluation are both objective and subjective. Moreover, this process entails students' perceptions. Recently, several researches have addressed this issue with varied methods for evaluating the teaching performance of a tutor. The majority of studies concentrate on strategies and theories of teaching performance evaluation, while very few papers look at quantitative analysis of teaching evaluation so as to improve tutor's performance. For example, Badri and Abdulla, (2004) examined higher education faculty performance including teaching, using AHP while Dong and Dai (2009) combined a fuzzy approach with neural networks using historical data that helps enhance teaching quality. Ramli et al (2010) proposed an approach of teaching performance evaluation with outlier data using a fuzzy approach. The above studies evaluate teaching performance objectively, but fail to design a scientific evaluation index system. In order to overcome this shortcoming, Chen et al (2015) introduced a fuzzy Analytic Hierarchy Process (AHP) approach for comprehensive teaching evaluation. While the fuzzy approach accommodates linguistic variables, the AHP (Saaty 1996) provides information on the importance of the criteria for evaluation. Kuzmanovic et al (2013) proposed a conjoint based approach, which accommodates the importance of criteria for student evaluation of teaching performance.

The evaluation of teaching performance must consider on one hand the importance of the criteria and on the other it must objectively assess the level of performance that can be expected from a teacher given the stage of their career and their experience. This will enable suitable targets of performance to be set for each teacher. There is to our knowledge no study that integrates these two aspects of teaching performance evaluation. The objective of this paper is to develop a method for teaching evaluation using SET questionnaires that integrates the importance of the criteria for evaluation from the student perspective and objectively derive targets for performance enhancement at teacher level. In this context this study uses the AHP method to derive the relative importance of each criterion of teaching performance, and DEA for deriving performance targets for each tutor. Although both the AHP and DEA have each been used extensively for higher education performance evaluation, the authors are not aware of any study that combines both the AHP and DEA to evaluate teaching performance in higher education.

The remainder of the paper is organised as follows: Section 2 describes the approaches to teacher evaluation in tertiary education. Section 3 outlines AHP and DEA as general purpose methods respectively for capturing decision maker preferences in multi criteria contexts and for performance evaluation and target setting. Section 4 introduces the proposed combined use of AHP and DEA for teaching evaluation.   Section 5 demonstrates the use of AHP to capture student preferences of a sample of students. Section 6 uses the preferences information and the feedback on teaching on a number of tutors to evaluate tutor performance. Section 7 concludes.

2. Approaches to Teacher Evaluation in Tertiary Education

Evaluation is applied in all levels of education with the area of higher education considered to be the most important. It is implemented in many aspects such as: evaluation of universities, evaluation of study programs (or faculties) and evaluation of academic staff, including teaching evaluation. In general, no matter which aspect we examine, building a robust evaluation framework is essentially a multiple criteria problem (Tsinidou et al, 2010). For that reason, in the literature, multiple criteria tools have often been used in the process. One of the reasons why teaching evaluation is complex is the fact that many of the criteria are qualitative in nature and there is the need of quantification. The choice of the quantification method as well as the scale used often raise questions for the objectivity of the process. Another important issue concerning the evaluation of higher education is the need to assign weights to the different evaluation criteria as it is logical that each factor does not contribute equally to the performance of the subject under evaluation (Marsh and Hocevar, 1991).

One of the most difficult aspects of higher education to be evaluated is that of academic staff. This is because it involves a large number of qualitative criteria which must be quantified in a rather objective manner. The evaluation of academic staff includes three dimensions: Research, Teaching and Community Services and major reasons for evaluations are to support decision making on rewarding, awarding or promoting as well as to evaluate teaching quality (Badri and Abdulla 2004; Marsh and Hocevar, 1991; Crumbley and Reichelt, 2009). Evaluation of teaching plays the important role of feedback both to teachers, in order to improve their performance (Marsh and Hocevar, 1991), (O'Hanlon and Mortensen, 1980) and to students in order to help them choose courses or supervisors (Crumbley and Reichelt, 2009).

The evaluation of research is carried out by the university authorities based on a combination of judgment and quantitative criteria such as number of publications, ranking or impact factor of journals where publications have appeared, number of conference participations etc. (Marsh and Roche, 1997; Crumbley and Reichelt, 2009). The evaluation of teaching is normally carried out using student questionnaires on content and delivery of courses, by peer evaluation or by self-evaluation by the tutors themselves (Marsh and Hocevar, 1991). In some cases (e.g. Greece) self-evaluations by academic departments which include teaching are submitted for external evaluation. This paper is concerned with the evaluation of teaching rather than research.

The instrument that is most frequently used for the evaluation of teachers is the SET (Student Evaluation of Teaching) questionnaire. The SET questionnaire is a means of controlling and/or measuring teachers' performance by the students (Crumbley et al, 2001). Two main issues concerning the SET questionnaires that are discussed in the literature are the dimensions and the criteria that should be included in these questionnaires, and the objectivity of the evaluations made by the students. Despite the fact that they are widely used, SET questionnaires are still an area of research debate. There is an ongoing discussion on the degree of correlation between teaching effectiveness and student learning (Uttl et al, 2016) as well as on potential bias in student evaluation (Badri et al, 2006) and background characteristics that are possibly influential to effective teaching (De Witte and Rogge, 2011).

Since teachers' evaluation is a multiple criteria process the proper choice and clear definition of criteria determines how fit for purpose is the questionnaire. The chosen criteria must have an important content and fully describe the subject under evaluation (Marsh and Roche, 1997). Various examples can be found in the literature of well defined sets of criteria, as shown in Table 1.

TABLE 1 NEAR HERE

There is a series of factors which can influence SET results returned by students. The factors can be grouped in four sets (Crumbley and Fliedner, 2002; Haladyna and Hess, 1994):

- **Student characteristics** such as disposition to instructor, gender, age, course level/year in school, graduate/undergraduate, expected grade, prior achievement (GPA), or personality;
- **Teaching conditions** such as class size, elective/required, discipline/department, work load/difficulty, course level, or time of day;
- **Instructor characteristics**, such as gender, academic rank, age, research productivity, or teaching level;
- **Content of the instrument**, uni-dimensional versus multi-dimensional; procedural factors such as purpose of ratings, anonymity of evaluators, presence of instructor, timing of administration, format, sampling, or leniency/severity.

The method proposed in this paper can capture through the AHP the effect of factors such as the above on student feedback. Evaluating student returns and teacher experience through DEA can lead to overall teaching performance evaluation and target setting as is elaborated in the next section.

3. Brief Outline of AHP and DEA

The combined use of AHP and DEA is suitable in cases where efficiency along with user preferences need to be taken into account (Yang and Kuo, 2003). In addition, it helps utilizing both qualitative and quantitative data. For example, the combined AHP and DEA approach has been adopted in internal audits of companies (Sueyoshi et al, 2009), and for the evaluation of sourcing firms that are undertaking the compiling of technical specifications of spare parts in the aerospace industry (Ferreira Filho et al, 2007).

To the authors' knowledge although SET and the AHP, and SET and DEA have been applied for performance evaluations of teachers, the integrated SET, AHP and DEA have not been previously used in teaching evaluation. This paper contributes to filling this knowledge gap by developing a method for teaching evaluation using students' responses of SET questionnaire, the AHP and DEA.

3.1 AHP

The AHP is a key method enabling the Decision Maker (DM) to provide preferences over criteria through pair-wise comparisons. Given two criteria *i* and *j* the DM is asked to return a value for (a_{ij}) in the form of a digit from 1 to 9 to reflect the degree to which *i* is preferred to *j* (if that is the case) or vice versa. The responses lead to the creation of a hierarchy matrix (A), for the relative importance between criteria *i* and *j* for $i > j$ i.e. *i* is preferred to *j*. The reciprocals are calculated for $i < j$ such that $a_{ji} = a_{ij}^{-1}$ while a_{ij} equals 1 for $i = j$. The data in the matrix is manipulated to derive relative weights for the criteria, and measures of the consistency of the decision maker preferences expressed. For more detail on how the AHP works the interested reader is referred to Saaty (1996).

3.2 DEA

DEA is a method for assessing the comparative performance of units setting a set of '*inputs*' against a corresponding set of '*outputs*'. Based on certain assumptions, the observed correspondences of inputs and outputs are used to construct potentially feasible, even if not observed, input-output correspondences. The *efficient* of these virtual or real correspondences are used as benchmarks to assess the relative efficiencies of the observed input-output correspondences. For a full introduction to DEA see Thanassoulis (2001).

There are numerous variants of DEA models depending on whether inputs or outputs are deemed controllable or part thereof, whether constant or variable returns to scale are assumed and so on. We shall use the generic DEA Model in (1) to assess teachers. The model assesses the efficiency of the input-output correspondence (x_{i0}, y_{r0}) by estimating the maximum factor φ by which the outputs y_{r0} could have been raised controlling for the levels of the inputs $x_{i0.}$

$$
\begin{aligned}
& Max\ \phi + \varepsilon \cdot \left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \right) \\
& s.t. \\
& \quad \sum_{j=1}^{n} \lambda_j \cdot x_{ij} + s_i^- = x_{io},\ i = 1,\dots,m \\
& \quad \sum_{j=1}^{n} \lambda_j \cdot y_{rj} - s_r^+ = y_{ro} \cdot \phi,\ r = 1,\dots,s \\
& \quad \sum_{j} \lambda_j = 1 \\
& \quad \lambda_j \geq 0,\ j = 1,..,n \\
& \quad s_i^- \geq 0,\ i = 1,..,m \\
& \quad s_r^+ \geq 0,\ r = 1,\dots,s
\end{aligned}
\tag{1}
$$

In model (1), x_{ij} are the level of input *i* while y_{rj} stand for the level of output *r* of DMU *j*; λ_j are decision variables whose values are determined by the model. Those with positive values identify the reference set (efficient peers) of DMU_0 under investigation. The slack variable that corresponds to input *i* is denoted s_i^- and the slack variable that corresponds to output *r* is denoted with s_r^+. It is assumed the inputs are m and the outputs s. The model assumes outputs are controllable and there are variable returns to scale between inputs and outputs. The efficiency of DMU_0 is $1/\phi^*$ where ϕ^* is the optimal value of ϕ in (1).

4. The proposed combined AHP and DEA method for teaching evaluation

Figure 1 shows schematically how the AHP and DEA are linked for the teaching evaluation proposed in this paper. The framework consists of ten steps. Step 1 is for conceptualizing the entire evaluation method in order to design the SET questionnaire, identifying the criteria for evaluation and selecting the most appropriate quantitative technique for deriving the index for performance. This requires the involvement of experts from multi-disciplinary perspectives, including teachers, students and institutional management. In step 2, the SET questionnaire is developed for gathering students' feedback on teaching performance in line with the objectives of the overall performance evaluation. Step 3 develops the AHP hierarchy through identification of criteria and sub-criteria in line with the SET questionnaire. Steps 2 and 3 are interrelated through the contents of the SET questionnaire and their links with the sub-criteria in the AHP hierarchy. Step 4 develops a DEA model for measuring tutors' efficiency with the identification of input and output variables. Step 4 is directly connected to step 3 in terms of relating the sub-criteria with the outputs of the DEA model. Step 5 is undertaking the SET survey and processing the data to derive tutors' performance against each question and overall performance. Step 6 is connected to step 3 as pair-wise comparison in criteria and sub-criteria levels, and subsequent synthesis of results lead to the overall importance of the sub-criteria. Step 7 combines the outcomes of step 5 i.e. SET survey results and step 6 (importance / weights of sub-criteria for teaching evaluation) to derive the weighted outputs for the DEA model. In step 8 the information for inputs is collected. In step 9 the DEA model is run to derive an efficiency evaluation of each tutor. In the last step the DEA results are analyzed to derive views about best practice in teaching at the collective teacher level, identify benchmark teachers, and set targets for improved performance by other teachers.

FIGURE 1 NEAR HERE

In this paper AHP is used to transform student declared preferences into quantifiable information (sub-criteria) which are then used within a DEA framework to arrive at a notion of 'efficiency' of each tutor in delivering service. Using the AHP, student preferences as captured on questions in the SET questionnaire are quantified. This information, along with other data about teacher prior experience, salary and research production are fed to DEA to calculate the scope for improvement in teaching for each teacher. The main aim of our approach is to address the following issues: Firstly, do students feel that all criteria used in the evaluation process are equally important? If not, then surely the relative weight of each criterion should be reflected in student evaluations. Secondly, if we have information on the relative weight students place on criteria it would be interesting to use it in comparing teachers' performance in light of the experience each teacher has at the current stage in their career. The basic intention of the paper is to derive the *assessment* of the teachers from students' perspectives. The functionality of the proposed model is tested using data from a Greek University.

5. Illustrative Application of the AHP to a set of Student Preferences

The proposed method has been applied to a set of realistic data from a higher education institution in Greece. The realistic data is derived from a set of real data using simulation as outlined later. The real data could not be used for reasons of confidentiality. The following sections demonstrate each one of the steps in Figure 1.

5.1 Case study sample

The sample for our experiment consisted of 120 students (we will call this the *aggregate sample*) from three courses at the University of Macedonia, in Thessaloniki, Greece. They have been classified in three separate groups of 30, 30 and 60 instances respectively. In particular:

- A sample of 30 responses was collected during a class of Operations Management (compulsory in one of the programs of the Business Administration Department). This course is taught at the 7th semester of an 8 semester degree program of studies.
- The second sample of 30 responses was collected during a course of Consumer Behaviour at the Economics Department. The course is offered by the Business Administration Department as an elective class to students of the Economics Department (5th semester).
- Finally, 60 responses were collected from a second compulsory course of the Business Administration Department, namely the Quantitative Analysis class of the 5th semester.

The selection of the courses was made in such a way that the three samples were mutually exclusive. In addition, the tutors of the courses sampled were different individuals so that students would not be biased by the image of one tutor, and furthermore, the aggregate sample contained both compulsory and elective courses.

5.2 SET Questionnaire for the case study

The format for the SET questionnaire (Appendix A) is designed in line with the institution's overall goal of the AHP and DEA based teaching performance evaluation method. The responses to this questionnaire are anonymous and questions were grouped into five sets namely, General Questions, Course Evaluation, Teacher Evaluation, Evaluation of Support Staff and Supporting Classes, and Other Questions. The evaluation is done with the use of a 5-point Likert scale except for the last group that has its own scale. (For the purposes of this paper the section on 'Other Questions' was deemed irrelevant and was dropped from the analysis.)

5.3 *Criteria and sub-criteria for the AHP Hierarchy*

The proposed AHP hierarchy consists of three levels – goal, criteria and sub-criteria as shown in Figure 2. The sub-criteria are in line with the SET questionnaire.

FIGURE 2 NEAR HERE

At the higher level the ultimate goal of the problem is placed which is the overall evaluation of teaching from the student perspective. At the next level down, there are two criteria – course and teacher. At this level we are interested in measuring the extent to which the students' satisfaction depends on the nature of the course itself or on the teacher's performance. At the lowest level the course is further analyzed through two sub-criteria: how interesting (overall) and how useful it is (usefulness refers to the career prospects fostered by the course as perceived by the student). In the original SET questionnaire these two sub-criteria are examined in one combined question, something which is misleading since the nature of the two criteria is significantly different. The teacher's dimension is analyzed in four sub-criteria: Preparation, Professionalism, Presence and Supporting Material. These sub-criteria are more complex and correspond to more than one question in the SET questionnaire. The following paragraphs provide the definitions of the sub-criteria as clearly explained to students before they completed the questionnaire:

Preparation: Preparation stands for the whole organization and presentation of the course (the course curriculum, the selection of the relative teaching material etc.) as well as the preparation of the teacher before each class.

Professionalism: Professionalism reflects the conduct of the teacher in terms of punctuality, access for students, timing of feedback and responsiveness to student requests.

Presence: This reflects the teacher's ability to communicate concepts to students, eagerness, encouraging participation, fostering questions and so on.

Supporting Material: Supporting Material comprises all the means that either help the teacher in delivering his/her lesson or accompany teaching, such as suggested literature, handouts, presentations used, exercises given, papers presented as well as supporting classes (delivered by teaching or technical assistants).

A second questionnaire is used to capture from the student perspective the relative importance of criteria. This was done through pair-wise comparisons of criteria using the AHP framework (Saaty, 1996) outlined earlier in section 3.1. The questionnaire had questions linked to the criteria reflected in Figure 2. For example Figure 3 shows the survey questions emanating from the 6 criteria of Level 2 of Figure 2. In respect of the first pairwise comparison of criteria in Figure 3 the student is asked: "*When evaluating the Teacher, what is the level of importance of his/her Preparation when compared to his/her Presence in clas*s ?" The student answers using the 9-point scale–e.g. the student scores say 7 at the appropriate place to indicate that Preparation by the teacher is significantly more important than his/her Presence in class. Clearly, all possible pairs of questions are asked at each hierarchical level (in total, n×(n-1)/2, where n denotes the number of criteria of the particular level of the hierarchy). The responses are then used within the AHP framework as outlined next.

FIGURE 3 NEAR HERE

5.4 Analysis and Results on Student Preferences

The 120 SET responses from our case study sample were processed through the *Expert Choice* software to derive the relative ranking of the criteria and sub-criteria from the student perspective. The *Expert Choice* software gives the user the opportunity firstly to structure, then prioritize the criteria and finally to calculate their weights based on pair-wise comparisons within the AHP method. The weights are either local (i.e. pertaining to a given priority level) or global, pertaining to overall preferences across all priority levels. The weights are normalised so that their absolute values are not significant, but their relative values are. We first present the results for the aggregate sample and then the results for the three separate sample groups.

a. Aggregate Sample

Table 2 presents the results of criteria and sub-criteria weights by Group and in aggregate. In the first column under Aggregate the local weights of sub-criteria are tabulated, the second column shows the global weights of each sub-criterion (derived by multiplying criteria weight with local weight of each sub-criterion), the third column is the ranking of the sub-criteria and the fourth column is the Consistency Ratio (*CR)* for each level. This ratio reflects the consistency of the preferences expressed by the respondents. It is recalled that in the AHP, an acceptable consistency ratio is below 0.1.

TABLE 2 NEAR HERE

The overall results reveal as the most important criterion of the hierarchy the tutor's Presence (G:0.268) followed by the course Usefulness (G:0.232) and Interest (G:0.171). It is interesting to note that what is more important for the students is how the tutor performs in class. It is also interesting that the other characteristics of the tutor are considerably less significant. In particular, the lack of Professionalism, which is often brought up by students as an area of dissatisfaction, is at least for this sample of students not so important. The consistency ratio is low enough to suggest that the AHP pair-wise comparisons are valid.

b. Sample group results and comparisons

The three sets of students noted above were: Group A: Business Administration, 7th semester (compulsory course), Group B: Economics 5th semester (course offered as elective from Business Administration Department), and Group C: Business Administration 5th semester (compulsory course). By contrasting the weights from the three different groups (Table 2) we observe some noticeable differences regarding the rankings of criteria. A difference that stands out is the fact that Groups A and C (both from Business Administration) consider the tutor as being much more important (L: 0.661 and L:0.623 respectively) than the Course itself (L:0.339 and L:0.377 respectively). On the other hand, for B, the result is the opposite since the Course (content) is thought as more important (L: 0.515) than the tutor (L: 0.485). This finding as noted below could be a consequence of the fact that Groups A and C were taking compulsory courses while Group B were taking an elective course. As far as the global weights are concerned, the results reveal that the criteria could be divided into two clusters: The first cluster contains the criteria Presence, Usefulness and Interest and the second the criteria Professionalism, Preparation and Supporting Material. The criteria of the first cluster are in all cases ranked in places 1-3 whereas, the criteria of the second cluster always have a rank ranging from 4 to 6. For the sake of simplicity let us call the first cluster of criteria as the primary and the second as the secondary criteria. For groups A and C the most important among the primary criteria is the tutor’s Presence (G:0.281 and G:0.278) followed by the Usefulness (G: 0.192 and G: 0.206) and third the overall Interest of the course (G:0.147 and G:0.172). For Sample B, Usefulness is the most important criterion (G:0.331), Presence follows (G:0.222) and finally Interest with (G:0.183). These results could be attributed to the fact that the course of group B is an elective one and therefore students aim at Usefulness rather than the tutor’s skills and expertise.

As for the secondary criteria, groups B and C rank them the same. In the fourth place we have Preparation (G: 0.160 and G: 0.097), fifth is the Supporting Material (G: 0.104 and G:0.090) and last is the Professionalism (G:0.081 and G:0.076). On the other hand, group A ranks as more important among the secondary criteria the Supporting Material (G:0.134), then follows Professionalism (G:0.125) and last is the Preparation (G:0.121). This discrepancy is attributed to seriousness and maturity in motivation as students become more independent in their study, strategies and goals. So in earlier semesters they consider as more important the Preparation of the teacher while senior students tend to be (and truly are) more demanding regarding the Professionalism of the teacher and need better Supporting Material.

The results of the above comparisons strongly suggest that final scoring for the teacher by SET questionnaire should take into account whether the course is compulsory or elective and the stage of study at which the students are.

c. Weighted Evaluations of Tutors through integrating SET responses with criteria importance

We can now derive the weighted performance of a tutor on each criterion. We use for this purpose the aggregate and global weights of criteria and sub-criteria from Table 2. Table 3 shows the weighted evaluations of two hypothetical (simulated) tutors as per their SET responses. The score for each question is the average of the subjects that the tutor teaches which have been collated from the student SET responses. Each SET question corresponds to one sub-criterion in the AHP hierarchy. The column of Table 3 headed “weight” is the importance of the criterion that corresponds to the specific question in SET, as derived using the AHP.

TABLE 3 NEAR HERE

The criteria included in the hierarchy may correspond to more than one question of the SET questionnaire. That is why some questions share the same weights (like questions 3&6, 5&11, 7&8 and 9&10). In those cases we have used the average of the two student evaluations to calculate the weighted mean of the corresponding criterion. Finally, to question 4 correspond two criteria: Interest and Usefulness. In this case the same evaluation is multiplied by two different weights.

The results of the weighted evaluation of two sample tutors are illustrated in Table 4. We would argue that evaluations of this type should be used when analysing student feedback on teaching. In fact, the idea is to use the AHP results from the three different samples and form different weights for integration in the SET evaluation process according to the course characteristics (e.g. compulsory – elective, early or late semester) of the degree programme.

TABLE 4 NEAR HERE

6. Application of the combined AHP and DEA approach to Teacher Evaluation

6.1 Deriving the comparative results on the performance of teachers

The weighted criteria values from the AHP for two tutors above are now expanded and used within a DEA model to illustrate how we can identify the scope for improvement on teaching for each tutor on each course they teach. We have used the real SET returns of tutors to derive a ‘realistic’ set of data to use within DEA. The use of realistic rather than actual data was necessary in order to keep confidential student feedback and tutor evaluation results. The realistic data have been generated by employing Monte Carlo simulations, using the real data distributions derived from student returns on each tutor. In order to reflect the real data as closely as possible, special care was taken during the simulations to include the full range of tutors by research output, higher or lower salary, years of experience etc. It is worth recalling that the study aim is not to assess tutors' performance per se but rather to demonstrate the proposed performance measurement method.

The tutors are regarded as Decision Making Units (DMUs) in the context of DEA. Tutors for the purposes of this illustration are seen as delivering two broad types of service, *Teaching* and *Research*. In the context of DEA these are *outputs*. Their attainments on teaching from the student perspective are captured in the manner outlined in the preceding section, using the AHP. Their attainments in research are reflected in research outputs such as refereed papers, contributed chapters etc. Attainments in teaching and research are set against two parameters that would reflect the expectation of attainment in teaching and research. These are *Salary* and *Experience* in an academic post which constitute *inputs* in the context of DEA. In essence, our model expects that the longer a tutor has been an academic and/or the higher his/her salary the better teaching quality and research outputs we should get, though not necessarily in a linear fashion, given that student satisfaction when expressed numerically has an upper bound and data is not scalable.

In the Greek University, normally the person teaching a course is also responsible for developing the course material. Assistant tutors draw from the tutor responsible for the course when assisting in the delivery of a course. Our assessment here refers to tutors responsible for delivering a course each. To the extent that student responses in SET questionnaires can be affected by context (see De Witte and Rogge, 2011) e.g. the type of course (science vs social science, the physical environment in which they are taught etc.) it is noted that our data relates to students taking courses in the same Department of a Business School, delivered in the same teaching facilities (Lecture theatres, seminar rooms etc.). Thus in effect contextual influences in the department are the same for all tutors.

Data for inputs and outputs of the DEA model are shown in Table 5. The output on Teaching is the AHP based *weighted evaluation* of the tutors and their course. We use within the DEA model the aggregate of "Course Interest + Course Usefulness + tutor's Professionalism + tutor's level of Preparation + tutor's Presence in class + Supporting Material" as one of our two outputs. A surrogate measure that was used for research output is the number of papers published by the tutor to date. Therefore, the final DEA model is:

- Inputs: Salary, Experience

- Outputs: Teaching ,Research

  Teaching was the aggregate weighted scores for: Course Interest + Course Usefulness + tutor's Professionalism + tutor's level of Preparation + tutor's Presence in class + Supporting Material

  Research was number of refereed publications.

Had the sample of tutors been larger, the 6 criteria "Interest" to "Supporting Material" in Table 5 could have featured as separate outputs within DEA permitting each evaluated tutor to allocate a different weight to each criterion (e.g. in the manner of the benefit of the doubt model (De Witte and Rogge, 2011)). This would have had the effect of permitting each tutor to choose the best weights for their mix of scores on the criteria to appear in the best possible light. The institution may in such a model impose weight restrictions to limit the flexibility tutors have in assigning weights to the criteria so that the resulting targets for performance by tutors would be more in line with the aims of the institution (e.g. see Rogge (2011)). Equally from the tutor perspective it could be argued that the two output components "Course Interest" and "Course Usefulness" are at best only indirectly affected by the tutor and have more to do with the prescribed syllabus for the course which may not be the tutor's sole responsibility. This would argue for perhaps treating these as a separate output with a weight restriction reflecting the lower influence a tutor has on these components compared to those reflected in the rest of the output. However, given the limited number of tutors and in order not to lose discriminatory power we have combined the scores on the six criteria into a single overall score of teaching attainment by the tutor to be used in the DEA model. Thus the model uses the AHP weights reflecting student perceptions to value aggregate teaching output. Through DEA, however, the model permits trade offs between teaching and research output to the best advantage of each tutor in turn.

TABLE 5 ABOUT HERE

Using the data in Table 5 we have solved the model in (1). The model assumes Variable Returns to Scale which is compatible with the non scale data pertaining to our inputs and outputs. We set within the DEA software (PIM of www.deasoftware.co.uk) the priority of the output 'Research' to zero so that the sole output with any weight was that of teaching. In this variant the model in (1) will yield an estimate of the maximum level a teacher's valuation by students could have been, controlling for research output, salary and years in post. We then take as a measure of '*efficiency*' the proportion a teacher's aggregate attainment on teaching represents of the maximum it could have been. Figure 4 shows the results.

FIGURE 4 ABOUT HERE

The picture at the aggregate level is quite good. The lowest efficiency is about 75% while the median is 94%. Some 33% of the sample are deemed benchmark in the sense that they represent the highest level of attainment in teaching evaluation by students compared to the rest of the faculty assessed. Clearly with 4 input-output variables we have limited discrimination on performance between teachers when the sample size is only 23 teachers. Nevertheless, for some 6 teachers, representing about 25% of the sample, we have already evidence that relative to the benchmark tutors on teaching their assessment by students can rise by over 20%. We return to the top and bottom on performance and how this information can be of use in the next section.

It is interesting to see whether it is the same teachers that are benchmark, or perform poorly on teaching for that matter, when both research and teaching are given equal priority to improve. The bars of Figure 4 show the efficiencies in teaching both when teaching is given sole priority, and when teaching and research are given equal priority to improve.

As can be seen the benchmark tutors (i.e. with 100% efficiency) are so whether teaching is given sole priority or teaching and research are given equal priority to improve. These would be good benchmarks for a department to seek for other tutors to emulate. In fact, all tutors have very similar efficiency in both scenarios on priority to improve. In very few occasions is the efficiency on teaching slightly higher when teaching and research are given equal priority. In those few cases a person's slightly higher efficiency in teaching when research is also prioritised to improve suggests he/she would have higher scope to improve in teaching if they diverted effort from research. However, the overwhelming picture in Figure 4 is that efficiencies in teaching are the same whether or not research is given equal to teaching priority to improve. In practice, this suggests for this sample of individuals poor attainment in teaching, such as there is, cannot be counter-balanced by attainment in research.

6.2 Using the findings on the comparative performance of teachers

The overarching aim of the analysis is to assess tutor performance on teaching as experienced and evaluated by students and guide such performance to improvement where necessary. Thus the results must be seen as an input towards the institution's formulation of a strategy to improve the way its students perceive teaching. This is not necessarily the same as assessing the overall performance of a teacher on teaching, though the approach outlined here can be part of a more rounded assessment of a teacher. The illustration in this paper is for the case where the DMUs are set up at course level, where the teacher has developed and teaches the course, in some cases with assistants who use his/hers teaching material. Where a teacher teaches on more than one courses the same teacher may be associated with two or more different DMUs within the assessment. Teacher-related multiple DMUs will differ only on teaching output as a given teacher may perform differently at different courses. This level of granularity of assessment would be more suitable where the aim is to guide a tutor to improved performance by course. If, on the other hand, the aim of the assessment is to capture the performance on teaching of a tutor across all courses he/she teaches, then the output on teaching used within the assessment model would need to be an aggregate of the assessment by the students of all the courses the teacher delivers. This tutor-level multi-course DMU may be more suitable for aims such as assessment of the tutor for promotion purposes but it is less useful for offering the tutor help in improving their teaching at course level.

We continue the illustration for the case where the assessment is at course level. We focus the illustration on the case where teaching is given sole priority to improve. This approach is in line with treating research output as a separable output from teaching in the sense that a tutor may not be permitted a poorer performance on say teaching on account of balancing better performance in research output (or vice versa).

As can be seen in Figure 4 we have in essence three categories of teachers on the way they are perceived by students. The *benchmarks* with efficiency of 100%, the *mid efficiency group* with efficiencies below 100% but above 80% and the '*low efficiency group*' teachers with efficiency below 80%. From the institutional perspective it would be interesting to see whether there is any feature that characterises each Group. Table 6 summarises the input-output data by efficiency Group.

TABLE 6 ABOUT HERE

The median values in Table 6 suggest that the attainment in teaching in absolute terms is not much different between teachers. Low efficiency tutors attain about 10% lower student evaluation than mid and benchmark tutors. Low efficiency tutors have been in post more than twice as long on average as benchmark teachers and they draw about 27% more on monthly salary than the benchmark tutors. These facts lead to an expectation derived from the DEA model that in fact on average they could raise their teaching evaluation by students by about 25%. This expectation may, however, be too optimistic. This is because there is an upper limit to the satisfaction rating (e.g. 5 on a Likert scale) that a student can award a tutor on a given criterion. If this is attained by a tutor after certain years in post, then additional years in post or salary for that matter, cannot lead to higher attainment in student evaluation.

Non proportionality between DEA model inputs and outputs of the foregoing type, and indeed the use of non scale data such as that from a Likert scale, are to an extent catered for by the use of a variable rather than constant returns to scale model. Nevertheless perhaps a more refined form of model (1) would be appropriate to reflect the capped nature of output data, e.g. through a non linear transformation of the years in post and salary prior to use within DEA. This form of refinement was not undertaken in this paper. Thus our findings need to be treated with caution, in the sense of indicating where potential for improvement exists rather than the precise scale of that potential.

It is noteworthy that research outputs are higher for the low in teaching efficiency individuals than for the benchmarks. However, the model solved has not permitted them to trade this off against poorer attainment in teaching. (Research was an exogenously fixed output in the DEA model solved.) As noted earlier, from the institutional perspective one aim would be to use the findings to guide individuals to improved performance in teaching. In this respect the focus should be on the individuals with low teaching efficiency. Each one of them will have been found inefficient relative to one or more particular benchmarks. The latter would be individuals from whom lessons on teaching can be drawn for the inefficient in teaching individual.

Table 7 shows the benchmarks used as comparators for each low teaching efficiency individual. The fractions under each benchmark show the weight the model gave each benchmark individual to arrive at attainment targets for each low teaching efficiency individual. It is clear that a key comparator has been DMU8. In the case of DMU19 DMU8 is the sole comparator. DMU8 carried the bulk of the weight also for the rest of the low efficiency individuals except for DMU6.

TABLE 7 ABOUT HERE

It is interesting to contrast the performance of DMU19 with that of DMU8 to see what lessons DMU19 may learn. Table 8 shows the raw data for the two DMUs. The two tutors have the same monthly salary and have been in post more or less the same duration – DMU8 for 22 and DMU19 for 26 years. However on all teaching components tutor 8 dominates tutor 19, and by a significant margin in proportional terms. The research output of tutor 8 is also significantly higher than that of tutor 19 and as can be seen in Figure 4 even when research can be given priority to improve, the efficiency of tutor 19 does not improve.

Even when a low efficiency tutor does not have a sole benchmark as tutor 19, a dominant benchmark can still prove useful as a role model for a corresponding low efficiency tutor. In this respect, Table 8 contrasts the performance of low efficiency tutor 23 with that of his/her dominant peer DMU8.

TABLE 8 ABOUT HERE

Tutors 23 and 8 have virtually the same duration in post and the same salary. However, tutor 8 dominates tutor 23 on all components of student evaluation, often by a considerable margin in proportional terms. Tutor 8 also offers a considerably better research outputs total.

In real life where the individuals concerned are in the same department the transfer of best practice in teaching from the benchmark tutors to low efficiency tutors should be sought by the institution. Indeed, in an assessment where a tutor may relate to more than one DMUs, the tutor's performance in one course can act as a role model for his/her performance in another course where the tutor has scope to improve.

7. Concluding remarks

Teaching evaluation, when used appropriately and findings are implemented can significantly enhance student experience. SET techniques have been developed to capture student perceptions of teaching quality and have been used to assess the performance of tutors. However, the SET returns need to be used with caution. Their analysis needs to reflect the varying preferences of students depending on the nature of their program of study, the stage in a student's education, whether a course is compulsory or optional and so on. This paper has addressed these issues in the context of teaching evaluation. The paper has also gone a step further and addressed the issue of what targets can be set in teaching for a tutor given the stage in their career and their performance in areas outside teaching, notably research.

The paper has developed in the form of Figure 1, a framework for designing SET questionnaires and analysing the returns using a combination of AHP and DEA. The SET allows us to capture student ratings for each tutor on the multiple criteria that characterize the delivery of a course by a teacher, by taking into account various evaluation factors. AHP is then used to capture the relative weights students would place in a given course on the various criteria characterizing that course. Using the SET ratings for each tutor and the criteria weights derived through AHP a set of weighted measures of attainment by each tutor are arrived at. The weighted measures can be weighted aggregates by category – e.g. Professionalism, Presence, Feedback Quality etc. pertaining to each tutor.

DEA is next engaged in order to place in the personal context of the tutor concerned, the teaching attainments. The DEA model estimates how much higher, if at all, could the attainments of the tutor have been when he or she is compared to other tutors controlling for the career stage of the tutor and their output in other areas, notably research. Clearly these are illustrative measures. They can be refined in a real application, for example by breaking down publications by the ranking of the journal in which they have appeared.

The paper suggests that the framework can be applied at varying levels of granularity depending on the aims of the assessment. If the aim is to simply offer tutors advice by course in which they each teach then each comparative unit of assessment would be a course – tutor combination. This offers the prospect of a tutor teaching more than one courses being a benchmark in one course and in need of better performance in another. If on the other hand the aim is to assess a tutor on teaching across all courses they teach the framework can be used by aggregating the ratings of a tutor across all courses they teach on. This approach will be less specific on advice about how a tutor may improve and what targets to aim at in each individual course they teach.

Using the approach in this paper the institution can identify at an aggregate level any features that may be common among the best attaining teachers. This type of information can in turn affect both how teachers are advised to improve their performance and also the recruitment policy of the institution in terms of features sought in candidates in future.

The paper illustrates the approach developed using sample data which are realistically close (simulation regenerated) to the exact data on tutors at a Greek university. The assessment reveals relatively little scope for further attainment in teaching by the tutors. This does not necessarily imply that there is no such scope. Rather most tutors perform similarly with the least well performing tutor having scope to raise their weighted attainment in teaching by about 25%. One third of teachers are benchmark performers. At the individual level the approach reveals one or more benchmark tutors each less efficient tutor can emulate to improve performance.

Clearly further enhancements to the approach developed here are possible. The SET questions can be honed better to each institution's criteria of good teaching. The research outputs of each tutor need to be captured more accurately and weighted for quality. The DEA model needs to perhaps be suitably modified to reflect the fact that the data used are right censored where rating in the form of a Likert scale is used. In addition, a point of interest for further research is how to account within teaching evaluations for contextual features such as class sizes, type of course (e.g. arts or science), location and size of teaching space etc. Such features are non-controllable from the tutor perspective yet they can influence student assessments of the tutor. One possible approach would be that put forth by De Witte and Rogge, (2011) where efficiencies are estimated conditional on the environmental factors which could influence student assessments of tutors.

In conclusion, the paper opens an avenue of research whereby using AHP and DEA in combination the teaching evaluations by students can be assessed in a manner that is more reflective of student preferences while teachers are set targets of attainment appropriate to the stage of their career.

References

1. Badri M and Abdulla M (2004). Awards of excellence in institutions of higher education: an AHP approach. International Journal of Educational Managemen 18(4): 224 - 242.

2. Badri M, Abdulla M, Kamali M and Dodeen H (2006). Identifying potential biasing variables in student evaluation of teaching in a newly accredited business program in the UAE. International Journal of Educational Management 20(1): 43 - 59.

3. Chen J, Hsieh H and Do Q H (2015). Evaluating teaching performance based on fuzzy AHP and comprehensive evaluation approach. Applied Soft Computing 28: 100 - 108.

4. Cherchye L, Moesen W, Rogge N and van Puyenbroeck T (2007). An introduction to "Benefit of the Doubt" composite indicators. Social Indicators Research 82(1): 111-45.

5. Crumbley D L and Fliedner E (2002). Accounting administrators' perceptions of student evaluation of teaching (SET) information. Quality Assurance in Education 10(4): 213–222.

6. Crumbley D L and Reichelt K J (2009). Teaching effectiveness, impression management, and dysfunctional behavior: Student evaluation of teaching control data. Quality Assurance in Education 17(4): 377–392.

7. Crumbley L, Henry B K and Kratchman S H (2001). Students' perceptions of the evaluation of college teaching. Quality Assurance in Education 9(4): 197–207.

8. De Witte K and Rogge N (2011). Accounting for exogenous influences in performance evaluations of teachers. Economics of Education Review 30(4): 641–653.

9. Dong P and Dai F (2009). Evaluation for teaching quality based on fuzzy neural network. Proceedings of the First International Workshop on Education Technology and Computer Science 1: 112–115.

10. Douglas J, Douglas A and Barnes B (2006). Measuring student satisfaction at a UK university. Quality Assurance in Education 14(3): 251 - 267.

11. DuBois P H and Garson E (1970). A history of psychological testing. Allyn and Bacon: Boston.

12. Ferreira Filho A J, Salomon V A and Marins F A (2007). Measuring the efficiency of outsourcing: an illustrative case study from the aerospace industry. In Loureiro G and Curran R (eds). Complex Systems Concurrent Engineering. Springer, pp 819–826.

13. Fitzpatrick J L, Sanders J R and Worthen B R (2004). Program evaluation: Alternative approaches and practical guidelines.

14. Frey P W, Leonard D W and Beatty W W (1975). Student ratings of instruction: Validation research. American Educational Research Journal 12(4): 435-444.

15. Gray M and Bergmann B R (2003). Student teaching evaluations: Inaccurate, demeaning, misused. Academe 89 (5): 44 - 46.

16. Haladyna T and Hess R K (1994). The detection and correction of bias in student ratings of instruction. Research in Higher Education 35(6): 669–687.

17. Hildebrand M (1971). Evaluating University Teaching.

18. Kuzmanovic M, Savic G, Gusavac B A, Makajic-Nikolic D and Panic B (2013). A Conjoint-based approach to student evaluations of teaching performance. Expert Systems with Applications 40(10): 4083-4089.

19. Guba E G and Lincoln Y S (1981). Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches. Jossey-Bass.

20. Marsh H W (1982). SEEQ: A Reliable, Valid, And Useful Instrument For Collecting Students'evaluations Of University Teaching. British Journal of Educational Psychology 52(1): 77–95.

21. Marsh H W (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. Journal of Educational Psychology 75(1): 150.

22. Marsh H W (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential baises, and utility. Journal of Educational Psychology 76(5): 707.

23. Marsh H W (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. International Journal of Educational Research 11(3): 253–388.

24. Marsh H W and Hocevar D (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. Teaching and Teacher Education 7(1): 9–18.

25. Marsh H W and Roche L A (1997). Making students’ evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. American Psychologist 52(11): 1187.

26. O’Hanlon J and Mortensen L (1980). Making teacher evaluation work. The Journal of Higher Education 664–672.

27. Rogge N (2011). Granting teachers the “benefit of the doubt” in performance evaluations. International Journal of Educational Management 25(6): 590-614.

28. Ramli N, Mohamad D and Sulaiman, N. H (2010). Evaluation of teaching performance with outliers data using fuzzy approach. Procedia-Social and Behavioral Sciences 8: 190-197.

29. Saaty T (1996). The Analytic Hierarchy Process, Volume 3 and 4. Pittsburgh: RWS Publishing.

30. Sueyoshi T, Shang J and Chiang W-C (2009). A decision support framework for internal audit prioritization in a rental car company: A combined use between DEA and AHP. European Journal of Operational Research 199(1): 219–231.

31. Thanassoulis E (2001). Introduction to the theory and application of data envelopment analysis. Massachusettes: Kluwer Academic Publishers.

32. Theofilidis C (1989). The Meta-Evaluation of the Evaluation of Programs (in Greek). Athens: New Education Publications.

33. Tsimboukis K (1979). Measurement and Evaluation in the Educational Sciences (in Greek). Athens: Orosimo Publications.

34. Tsinidou M, Gerogiannis V and Fitsilis P (2010). Evaluation of the factors that determine quality in higher education: an empirical study. Quality Assurance in Education 18(3): 227 - 244.

35. Uttl B, White C A and Gonzalez D W (2016). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. Studies in Educational Evaluation.

36. Warrington W G (1973). Student evaluation of instruction at Michigan State University. In Proceedings: The first invitational conference on faculty effectiveness as evaluated by students 164–182.

37. Worthen B R and Sanders J R (1987). Educational evaluation: Alternative approaches and practical guidelines. Longman Pub Group: New York.

38. Yang T and Kuo C (2003). A hierarchical AHP/DEA methodology for the facilities layout design problem. European Journal of Operational Research 147(1): 128–136.

Appendix A. The SET questionnaire

A usual 5-point Likert scale is used by the student to express his/her level of agreement with the statement.

Group 1 General statements

1. The overall performance of the teacher was good.
2. The quality of the course was high.

Group 2 Course Evaluation

3. The organization and the presentation of the course were complete.
4. The subject of the course was interesting and useful for your studies.
5. The course material (books, handouts, slides, exercises, papers etc) were satisfactory for the course needs.

Group 3 Teacher Evaluation

6. The tutor was well prepared for the class.
7. The tutor had good transmissibility.
8. The tutor encouraged questions and in general the participation in class.
9. Whenever I needed to meet the tutor for discussing questions or problems he/she was there during his/hers office hours.
10. The tutor was punctual for the classes.

Group 4 Evaluation of Supportive Classes and Supportive Teaching Staff (to be answered only if supportive classes exist).

11. The quality of the supportive classes was high.
12. The overall performance of the supportive teaching staff was good.

Group 5 Other Questions (specific scales)

13. Classes attend frequency (not obligatory attendance)

    1=not at all, 2=rarely, 3=often, 4=very often, 5=always

14. According to your experience with other courses you would characterise this course as:

    1=very easy, 2=easy, 3=of average difficulty, 4=difficult, 5= very difficult

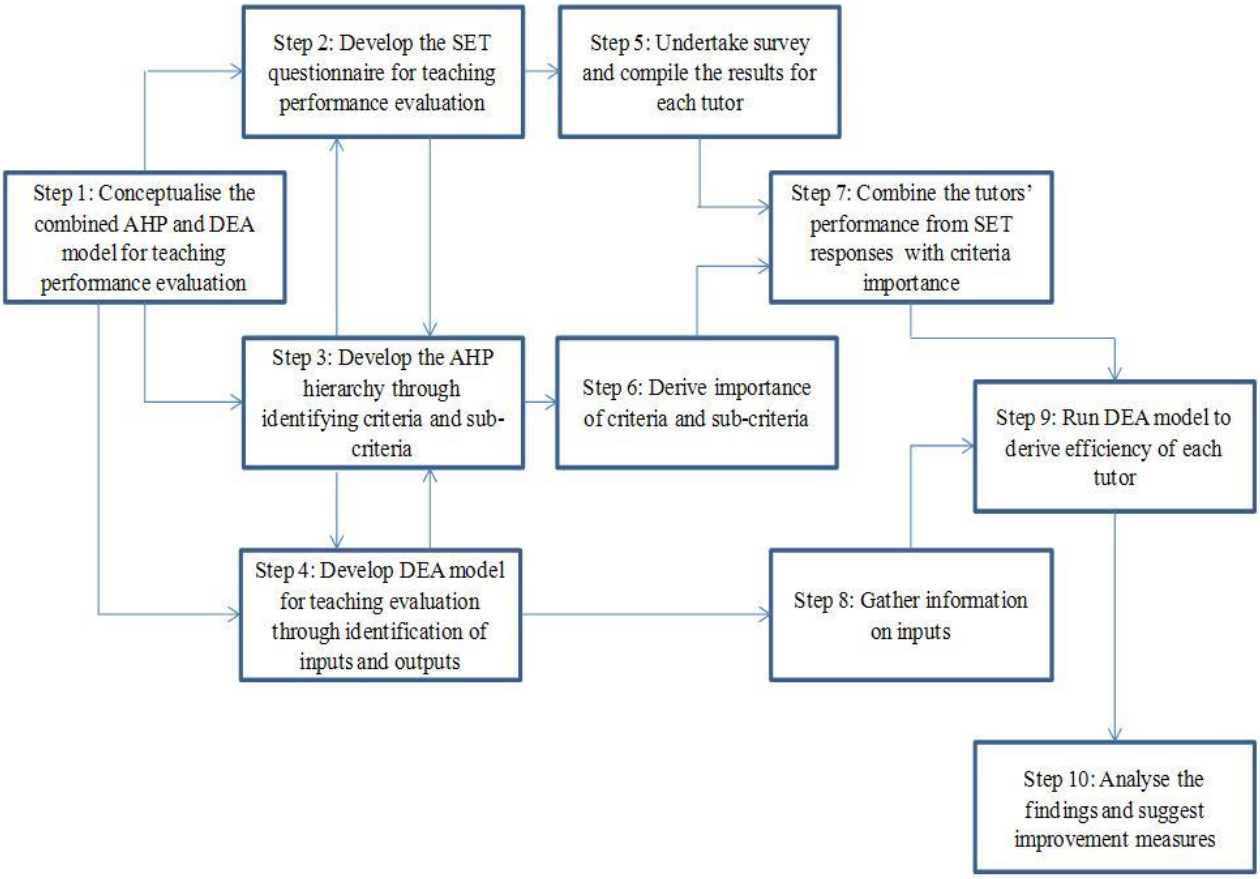
Step 1: Conceptualise the combined AHP and DEA model for teaching performance evaluation
Step 2: Develop the SET questionnaire for teaching performance evaluation
Step 3: Develop the AHP hierarchy through identifying criteria and sub-criteria
Step 4: Develop DEA model for teaching evaluation through identification of inputs and outputs
Step 5: Undertake survey and compile the results for each tutor
Step 6: Derive importance of criteria and sub-criteria
Step 7: Combine the tutors' performance from SET responses with criteria importance
Step 8: Gather information on inputs
Step 9: Run DEA model to derive efficiency of each tutor
Step 10: Analyse the findings and suggest improvement measures

Figure 1 Proposed combined SET, the AHP and DEA framework for teaching performance evaluation

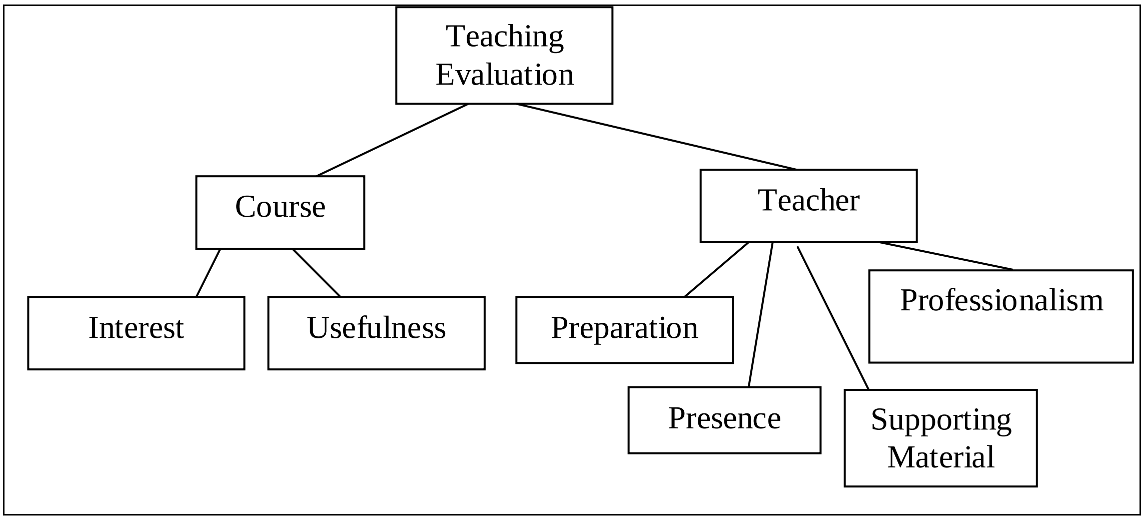

Figure 2. The AHP hierarchy model

Pair-wise Comparisons of Tutor Characteristics																		
Preparation	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	**Presence**
Preparation	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	**Professionalism**
Preparation	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	**Supporting Material**
Presence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	**Professionalism**
Presence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	**Supporting Material**
Professionalism	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	**Supporting Material**

Please compare on importance from your point of view each pair of tutor characteristics by circling the appropriate number on the scale.
1: Equal importance, 2-3: Moderate importance, 4-5: Strong importance, 6-7: Very strong importance, 8-9: Extreme importance

Figure 3: Pairwise Comparison of Criteria at Level 2 of Figure 2

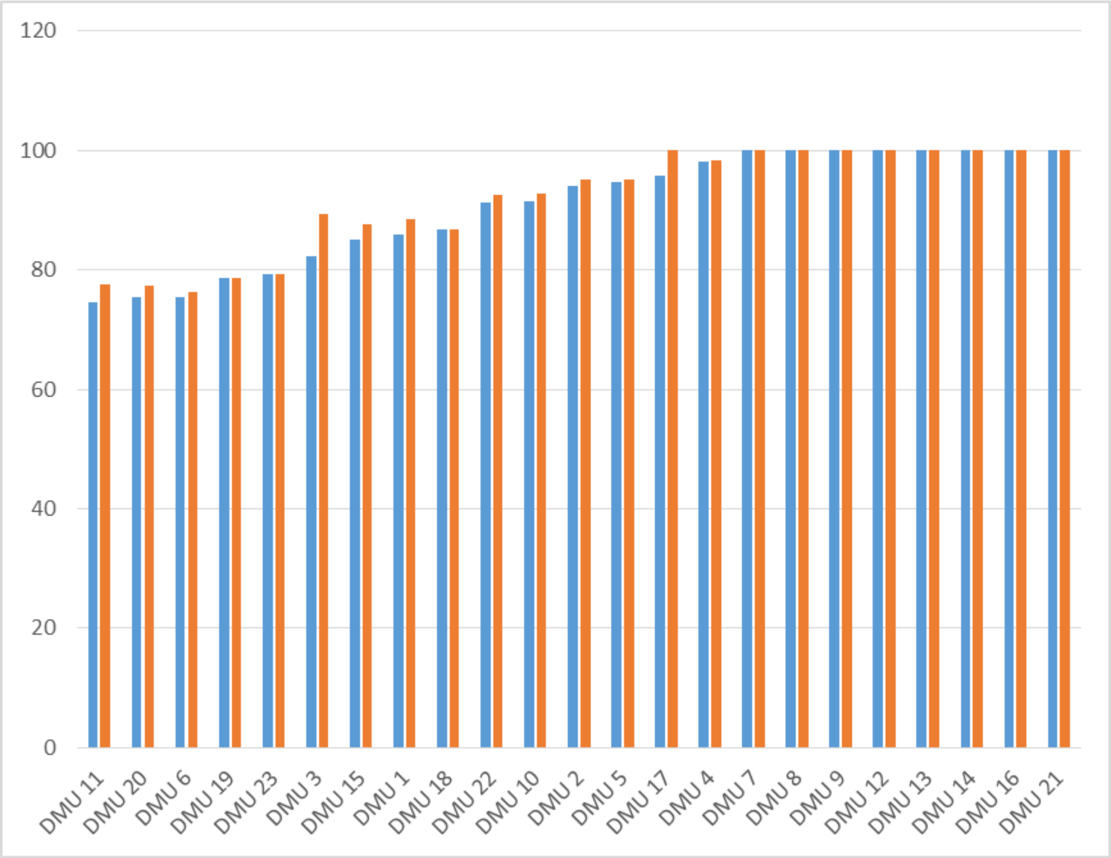

Figure 4: Teaching evaluation % of maximum it could have been. (Left bars when teaching is given sole priority to improve, right bars for teaching and research equal priority to improve.)

Table 1. Criteria typically used to evaluate teaching in universities

SET method	Reference	Criteria
Frey's Endeavor instrument, Student Description of Teaching (SDT) questionnaire	Frey et al, 1975;Marsh, 1982;Marsh, 1987; Hildebrand et al, 1971	Presentation Clarity, Workload, Personal Attention, Class Discussion, Organization/ Planning, Grading, and Student Accomplishments
Marsh's Student Evaluations of Educational Quality (SEEQ) instrument	Marsh, 1982;Marsh, 1983; Marsh, 1984;Marsh, 1987	Learning/Value, Instructor Enthusiasm, Organization/Clarity, Individual Rapport, Group Interaction, Breadth of Coverage, Examinations/Grading, Assignments/Readings
The Michigan State SIRS instrument	Warrington, 1973	Learning/Value, Instructor Enthusiasm, Organization/Clarity, Individual Rapport, Group Interaction, Breadth of Coverage, Examinations/Grading, Assignments/Readings

Table 2.Weights of the aggregate and three different sample groups

	Aggregate Sample			Group A (compulsory)			Group B (elective)			Group C (compulsory)		
Teaching Evaluation Criteria	Local	Global	Rank	Local	Global	Rank	Local	Global	Rank	Local	Global	Rank
1. Course	L: 0.403			L:0.339			L:0.515			L:0.377		
1.1. Interest	L: 0.424	G: 0.171*	3	L:0.433	G:0.147	3	L:0.356	G:0.183	3	L:0.455	G:0.172	3
1.2. Usefulness	L: 0.576	G: 0.232	2	L:0.567	G:0.192	2	L:0.644	G:0.331	1	L:0.545	G:0.206	2
2. tutor	L: 0.597			L:0.661			L:0.485			L:0.623		
2.1. Professionalism	L: 0.150	G: 0.089	6	L:0.189	G:0.125	5	L:0.156	G:0.076	6	L:0.129	G:0.081	6
2.2. Preparation	L: 0.191	G: 0.114	5	L:0.183	G:0.121	6	L:0.200	G:0.097	4	L:0.257	G:0.160	4
2.3. Presence	L: 0.449	G: 0.268	1	L:0.425	G:0.281	1	L:0.458	G:0.222	2	L:0.447	G:0.278	1
2.4.Supporting Material	L: 0.210	G: 0.125	4	L:0.202	G:0.134	4	L:0.186	G:0.090	5	L:0.167	G:0.104	5
Overall Consistency	0.01			0.02			0.04			0.01		

*Global weights of sub-criteria = local weight × criteria weight; E.g. 0.171 = 0.424 × 0.403

Table 3. SET statements for tutors' evaluations.

SET Statements	Mean Evaluation		Weight*
	tutor 1	tutor 2	
3 The organization and the presentation of the course were complete	4.40	4.02	0.114
4 The subject of the course was interesting and useful for your studies	4.37	4.50	0.171/0.232
5 The course material (books, handouts, slides, exercises, papers etc) were satisfactory for the course needs	4.05	3.80	0.125[a]
6 The tutor was well prepared for the class	4.22	4.10	0.114
7 The tutor had good transmissibility	3.06	4.02	0.268[b]
8 The tutor encouraged questions and in general the participation in class	4.50	4.14	0.268[b]
9 Whenever I needed to meet the tutor for discussing questions or problems he/she was there during his/hers office hours	4.28	3.98	0.089[c]
10 The tutor was punctual for the classes	4.19	4.00	0.089[c]
11 The quality of the supportive classes was high	4.32	4.10	0.125[a]

*Refer Table 2, aggregate sample global weights column – Identical weights with superscripts correspond to the same criterion.

Table 4.Weighted evaluations of tutors

Teaching Evaluation Criteria	Corresponding Questions	Mean Evaluation		Weight	Weighted Mean	
		tutor 1	tutor 2		tutor 1	tutor 2
1.1. Interest	4	4.37	4.50	0.171	0.747	0.770
1.2. Usefulness	4	4.37	4.50	0.232	1.014	1.044
2.1. Professionalism	9&10	4.235	3.99	0.089	0.377	0.355
2.2. Preparation	3&6	4.31	4.06	0.114	0.491	0.463
2.3. Presence	7&8	3.78	4.08	0.268	1.013	1.093
2.4.Supporting Material	5&11	4.185	3.95	0.125	0.523	0.494

Table 5. Inputs and outputs of the DEA phase of the model

	Outputs								Inputs	
	Interest (Y1)	Usefulness (Y2)	Professionalism (Y3)	Preparation (Y4)	Presence (Y5)	Supporting Material (Y6)	Overall evaluation score (Y1+Y2+Y3+Y4+Y5+Y6)	Research Work	Salary	Experience
tutor 1	0.7470	1.0140	0.3770	0.4910	1.0130	0.5230	4.165	26	4100	15
tutor 2	0.7700	1.0440	0.3550	0.4630	1.0930	0.4940	4.219	24	3450	10
tutor 3	0.6940	0.9410	0.3600	0.4330	1.0050	0.5250	3.958	31	4100	17
tutor 4	0.8020	1.0881	0.4272	0.5535	1.3078	0.5469	4.7255	18	3450	12
tutor 5	0.7097	0.9628	0.3702	0.4737	1.0358	0.4875	4.0397	10	3000	5
tutor 6	0.6088	0.8259	0.2866	0.3939	0.7826	0.3875	3.2853	9	3000	6
tutor 7	0.7984	1.2150	0.4054	0.5521	1.4223	0.5527	4.9459	11	3450	9
tutor 8	0.8565	1.0957	0.4345	0.5754	1.5353	0.5984	5.0958	34	4100	22
tutor 9	0.6652	0.9025	0.4014	0.4893	0.9943	0.4541	3.9068	5	3000	2
tutor 10	0.7456	1.0115	0.3533	0.4856	1.1082	0.5158	4.22	22	3450	11
tutor 11	0.6413	0.9440	0.3418	0.3825	0.8951	0.5581	3.7628	25	4100	18
tutor 12	0.7182	0.9744	0.3467	0.4931	0.9970	0.5650	4.0944	9	3000	3
tutor 13	0.6874	0.9326	0.3235	0.4190	0.8857	0.5438	3.792	38	4100	16
tutor 14	0.7097	0.9628	0.3596	0.5273	1.0747	0.5288	4.1629	31	3450	9
tutor 15	0.8037	1.0904	0.3987	0.4714	1.0519	0.4788	4.2949	27	4100	18
tutor 16	0.7280	0.9880	0.3850	0.4680	1.0880	0.5700	4.227	21	3000	8
tutor 17	0.6900	0.8960	0.3910	0.4280	1.0550	0.5310	3.991	31	3450	12
tutor 18	0.7920	1.0740	0.3860	0.4820	1.1480	0.4970	4.379	13	4100	18
tutor 19	0.7350	0.9560	0.3250	0.4350	1.0520	0.5060	4.009	25	4100	26
tutor 20	0.6990	0.9490	0.3680	0.4290	0.8540	0.5290	3.828	26	4100	20
tutor 21	0.8140	0.8820	0.3910	0.5090	1.1500	0.4400	4.186	13	3450	9
tutor 22	0.7400	1.0000	0.3820	0.5070	1.2060	0.5260	4.361	18	3450	11
tutor 23	0.6380	1.0120	0.3630	0.4460	1.1000	0.4670	4.026	22	4100	21

Table 6. Median values by Group on efficiency

Group on Teaching efficiency	Salary (scaled)	Years in post	Research outputs	Teaching evaluation	Teaching Efficiency	Number of DMUs
Low	4100	20	25	3.828	75.54	5
Mid	3450	12	23	4.219	91.41	10
Benchmark	3225	9	17	4.1744	100	8

Table 7. Correspondences between the 5 low efficiency and benchmark individuals.

	Benchmarks			
	DMU7	DMU8	DMU12	DMU21
DMU 6	0.28	0	0.5	0.23
DMU 11	0.31	0.69	0	0
DMU 19	0	1	0	0
DMU 20	0.15	0.85	0	0
DMU 23	0.08	0.92	0	0

Table 8. Contrasting low efficiency tutors 19 and 23 with benchmark tutor 8

Tutor	Interest (Y1)	Usefulness (Y2)	Professionalism (Y3)	Preparation (Y4)	Presence (Y5)	Supporting Material (Y6)	Overall evaluation score (Y1+Y2+Y3+Y4+Y5+Y6)	Research Output	Salary	Experience
8	0.857	1.096	0.435	0.575	1.535	0.598	5.096	34	4100	22
19	0.735	0.956	0.325	0.435	1.052	0.506	4.009	25	4100	26
23	0.638	1.012	0.363	0.446	1.100	0.467	4.026	22	4100	21