ASSESSING RESEARCH EFFECTIVENESS: A COMPARISON OF ALTERNATIVE NON-PARAMETRIC MODELS

Giannis Karagiannis[1] and Georgia Paschalidou[2]

[1] Professor, Department of Economics, University of Macedonia, Thessaloniki, Greece; email: karagian@uom.gr; tel. ++30 2310 891 759 (corresponding author)

[2] Ph.D student, Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece

ABSTRACT: In this paper we examine three alternative *a posteriori* weighting schemes with variable, common and restricted weights in order to assess research productivity by means of two seemingly similar non-parametric models: the BoD and the K&H model. Our empirical results, based on different types of faculty members' publications, show that there is more variability in the estimated effectiveness scores among alternative weighting schemes within each model rather than between models for any particular weighting scheme. In addition, we also found that the effectiveness scores from the BoD model are greater than or equal to those from the K&H model for the variable- and the restricted-weights schemes while there is no clear pattern between the BoD and the K&H effectiveness scores from the common-weights scheme.

ACKNOWLEDGMENT: We would like to thank two anonymous referees and the associate editors for helpful comment and suggestions. A previous version of this paper was presented at 2nd Workshop on Education Efficiency hold in London, Sept. 26-29, 2014 and at the 2014 Asia-Pacific Productivity Conference hold in Brisbane, July 8-11, 2014. This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) – Research Funding Program: THALES-Investing in Knowledge Society through the European Social Fund.

October 2016

ASSESSING RESEARCH EFFECTIVENESS: A COMPARISON OF ALTERNATIVE NON-PARAMETRIC MODELS

1. Introduction

Three are the main features in assessing research productivity: *first*, which aspects of performance to be considered; *second*, how to appropriately measure them and *third*, what weights to assign to each one. Production (volume), impact and excellence (quality) are the aspects of research performance usually considered. The number and type of publications are by far the most common indicators used to quantify research output, coupled with relevant quality attributes (e.g., journal impact factors) and recognition measures proxied either by citation counts or commonly used bilbiometric indicators such as the h- or the g-index. On the other hand, we may distinguish between *a priori* and *a posteriori* weights. The former are set prior to the evaluation process and usually come from experts' or stakeholders' opinions. They are common to all evaluated decision making units (DMUs) and in most of the cases, assign equal weights to all sub-indicators. In contrast, *a posteriori* weights are derived during the evaluation process by means of an optimization procedure. They may be variable or common to the evaluated DMUs and may or may not reflect experts' and stakeholders' opinions.

The variable-weights scheme is based on the idea that it may be better each evaluated DMU to derive its own weights (under certain regulatory conditions) in order the evaluator(s) to avoid complains about its final ranking afterwards. Data envelopment analysis (DEA) is the main operational tool behind the variable-weights scheme. Some researchers (e.g., Kao and Hung, 2005; Wang, Luo and Lin, 2011) have argued however that comparison and ranking of units are meaningful only when they are conducted on common grounds and thus they favor the use of common but not necessarily equal weights, which though are not determined *a priori*. Several variants of DEA (see e.g. Kao and Hung, 2005) have been developed for this purpose.

In addition, we may consider a combination of *a priori* and *a posteriori* weights. In this case, two rather distinct approaches may be used: peer appraisal by means of cross efficiencies and DEA with weight restrictions. In the former case, the value norms of all evaluated DMUs are taken into account when assessing the research performance of each DMU while in the latter case, the weights assigned to sub-indicators are restricted to lie between pre-specified ranges determined by experts or stakeholders.

In this paper we compare and contrast these three *a posteriori* weighting schemes in order to assess research performance by means of two seemingly similar non-parametric models that share a common feature: they account only for the output side of the research process, acting actually as an aggregator function of the relevant sub-indicators. In that sense they measure effectiveness rather than efficiency of research or as Doyle and Green (1994a) put it they examine research excellence rather than value for money. These two models are the Benefit-of-the-Doubt (BoD) and the Kao and Hung (2003) (K&H hereafter) model.[1] For the BoD model, it turns out that the three alternative *a posteriori* weighting schemes with variable, common and restricted weights correspond to three different aspects of performance evaluation based respectively on self-appraisal, peer-appraisal and value judgment. In the case of the K&H model, on the other hand, three alternative *a posteriori* weighting schemes correspond respectively to the conventional formulation of the model, the Kao and Hung (2005) compromise solution approach, and the partial average formulation of Ng (2007, 2008).

Some of these variants of the BoD and the K&H models have been previously used to assess research effectiveness but to the best of our knowledge there is no comparative study for all of them using the same data set.[2] In particular, Kao and Pao (2009) employed the conventional and the common-weights K&H models to evaluate research productivity among management faculty members in Taiwan. de Witte and Rogge (2010) relied on the conventional and the restricted-weights BoD models to estimate research effectiveness among researchers at the department of Business Administration of the Hogeschool Universiteit Brussel. Kao, Liu and Pao (2012) used the conventional BoD model to assess research performance of Taiwanese professors in three areas of management, namely management information systems, production and operations management, and marketing.[3] Our comparative study is

based on the publication records of the faculty members in the department of Economics at the University of Macedonia, Greece during the period 2000-2006.

The rest of this paper proceeds as follows: in the next section we present and compare the BoD and the K&H models under the three alternative weighting schemes. In the third section we describe our data set and the definitions of research sub-indicators considered. Our comparative empirical results are reported in the fourth section of the paper. Concluding remarks follow in the last section.

2. Non-parametric Models for Assessing Research Effectiveness

The main purpose of using DEA to assess research productivity is to obtain, through an optimization procedure based on linear programming, *a posteriori* weights to aggregate research inputs and outputs in order to derived a single metric, by means of an evaluation score reflecting relative achievement.[4] The *a posteriori* weights may be variable (i.e., DMU-specific) or common, and may or may not reflect (at least partially) experts' or stakeholders' opinions.

The variable weights resulting from conventional DEA models reflect its underlying assumption that each evaluated DMU is allowed to choose, under certain regulatory conditions, its own set of input and output weights in order to show in the best possible light relative to other DMUs in the sample. It is thus able to exaggerate its own advantages and at the same time to downplay its own weaknesses in order to obtain the maximal possible evaluation score. Thus the variable-weights scheme expresses in the best possible way the interest of the evaluated DMUs (faculty members in our case) who may assign extremely low or high weights to certain output(s) that through improper incentive mechanisms may result in undesirable specialization of research. With this weighting scheme, neither internal management (e.g., department's head, school's dean, etc.) nor other faculty members have a "say" on the research performance and priorities of each faculty member even in the same institution. However, the variable-weights scheme is the most affirmative in its resulting outcomes: if afterwards someone is still weak relative to other DMUs in the sample this cannot be put down to the choice of input and output weights.

On the other hand, some authors (e.g., Kao and Hung (2005) and Wang, Luo and Lin (2011)) have argued that variable weights are useful only for classifying the evaluated DMUs as efficient or inefficient but not for ranking. In contrast, they suggest that comparison and ranking is meaningful only if it is conducted on common

grounds and thus they favor the use of common but not necessarily equal weights.[5] This gives the impression of a more fair treatment since the relative importance of each research output is the same for all evaluated DMUs and consequently, there are no incentives for undesirable specialization. Besides this, it turns out that the use of the common-weights scheme in the BoD model results in average cross efficiency scores (see Karagiannis and Paleologou, 2014). This is of particular interest as it shed some light on the managerial implication of using common weights in assessing research productivity: the notion of cross efficiency provides the basis of giving the right to every faculty member to have a "say" about the performance of other faculty members in the same institution (Oral *et al.*, 2014). The value norms (i.e., the DEA weights) of all evaluated units are taken into account when assessing the performance of each faculty member. Nevertheless, internal management (e.g., department's head, school's dean, etc.) still does not have a "say" on the research performance and priorities of faculty members and the weights are obtained by averaging model-based weights across the evaluated DMUs.

Lastly, a combination of *a posteriori* and *a priori* weights (i.e., model- and experts/stakeholders-based) is also possible. This requires the use of what is called value judgment DEA where the weights assigned to (some or all) inputs and outputs are constrained to satisfy *a priori* restrictions in order to eliminate the possibility of assigning zero values to particular inputs and/or outputs and more generally, to ensure that the DEA weights accord with intuition and/or research priorities. In the case of research performance evaluation, these *a priori* restrictions may reflect the opinions of external evaluation committees, the government, stakeholders, internal management (e.g., department's head, school's dean), or even of faculty members. Then, the resulting model-based weights are still variable across DMUs but their estimated values are restricted to satisfy certain prior restrictions regarding the importance of each research output and thus do not allow for extreme specialization.

The second methodological aspect to be considered at the outset of the evaluation process is related to the aggregation level at which the assessment exercise will be conducted. The level of aggregation runs from the individuals (i.e., faculty members) to institutions, namely departments, schools/colleges, and the university as a whole (Doyle and Green, 1994a). We can thus evaluate the research productivity of faculty members as well as that of the departments or the universities that belong to. In this paper, we start from the research productivity of faculty members and then we

propose an appropriate way to aggregate them to the institution level of interest. This requires a consistent way to aggregate efficiency and effectiveness scores from the individual to the institution level.

The third methodological aspect to be considered is whether resources related to research activities will be taken into account or not. In a sense this is may be viewed as distinguishing between efficiency (benefits realized versus resources utilized) and effectiveness (ability to state and achieve goals), when in the latter targets (goals) are determined by observed behavior (i.e., best practice) and our objective is to evaluate the extent to which they are achieved (Prieto and Zofio, 2001). Then, efficiency compares the outcome(s) of the research related activities relative to the resources employed for this purpose (i.e., value for money) while effectiveness compares only the outcome(s) of the research related activities (i.e., research excellence) and not the means to achieve them. Conventional DEA models are used to estimate efficiency of research activities while the pure-output DEA model (i.e., a model without explicit inputs) is used to estimate research effectiveness. In such a setting it is assumed that the faculty members have roughly equal research opportunities and resources at their disposal and thus one can disregard them in evaluating research performance. Lovell and Pastor (1999) have shown however that the pure-output DEA model is equivalent to a DEA model with a single constant input. This normalization in the input side is compatible with Koopmans' idea of a helmsman who has at his/her disposal a unitary quantity of an aggregate input and attempts to steer all of research outputs towards their maximum levels.

2.1 *Variable Weights*

The BoD model is essentially a tool for aggregating linearly quantitative performance sub-indicators into a single composite indicator when the exact weights are not known *a priori* (Cherchye *et al.*, 2007). For each evaluated DMU, this is done by implicitly assigning less (more) weight to those sub-indicators or aspects of performance that the particular DMU is relatively weak (strong) compared to all other DMUs in the sample. As a result, the estimated weights are allowed to vary across DMUs and time.

In technical terms, the BoD is a benchmarking model that has a DEA-type structure in the sense that the composite indicator is defined by the ratio of an actual to a benchmark measure, both of which are given by the weighted sum of the

analyzed sub-indicators. Since the composite indicator is designed to take values in the [0,1] interval, benchmark performance is indicated by the value of one (Cherchye *et al.*, 2007). In determining overall performance, the weights are selected in such a way as to maximize the value of the composite indicator of the evaluated DMU. This in turn guarantees that any other weighting scheme would worsen this DMU's efficiency score. Moreover, when these weights are used by any other DMU in the sample they would not result to a composite indicator greater than one. The resulting weights are determined endogenously by solving for each evaluated DMU the following problem:

$$I^k = \max_{s_i^k} \sum_{i=1}^{N} s_i^k I_i^k$$

$$st \sum_{i=1}^{N} s_i^k I_i^j \leq 1^j \quad \forall\, j = 1, \ldots, K \qquad (1)$$

$$s_i^k \geq 0 \qquad \forall\, i = 1, \ldots, N$$

where I_i^k is the i[th] sub-indicator of the k[th] DMU, s_i^k are the weights to be estimated, *j* is used to index DMUs and *i* to index sub-indicators which in our case correspond to different research outputs.

The BoD model is equivalent to the multiplier form of the Charnes, Cooper and Rhodes (1978) input-oriented, constant-returns-to-scale DEA model when there is a single constant input that takes the value of one for all evaluated DMUs.[6] Based on this, the dual formulation of the BoD model is given as:

$$I^k = \min_{\lambda_j^k} \sum_{j=1}^{K} \lambda_j^k 1^j$$

$$st \sum_{j=1}^{K} \lambda_j^k I_i^j \geq I_i^k \quad \forall\, i = 1, \ldots, N \qquad (2)$$

$$\lambda_j^k \geq 0 \qquad \forall\, j = 1, \ldots, K$$

where λ refers to intensity variables. This implies that the value of the composite indicator is in fact equal to the sum of the intensity variables. From the inequality constraint in (1) it is clear that the BoD model exhibits constant returns to scale.[7]

On the other hand, the K&H model has a similar structure in the sense of deriving a set of *a posteriori* weights that maximize the value of a composite indicator but now under the assumption that, for each evaluated DMU, this set of weights satisfies an adding-up/normalization constraint. The K&H model is written as:

$$\begin{aligned} E^k &= \max_{u_i^k} \sum_{i=1}^{N} u_i^k I_i^k \\ st &\sum_{i=1}^{N} u_i^k = 1 \qquad (3) \\ & u_i^k \geq 0 \qquad \forall\, i = 1, \ldots, N \end{aligned}$$

Even though the two models have the same objective function they differ in terms of the underlying constraints, which in the case of the K&H model render a linear programming rather a DEA-type structure. In addition, in the K&H model there is only one (equality) constraint, besides the non-negativity constraints on the weights, while in the BoD model the number of (inequality) constraints is equal to the number of evaluated DMUs.

Besides these differences, Kao *et al.* (2008) have shown that the two models are related to each other as long as the set of sub-indicators to be aggregated are normalized at the outset to lie in the [0,1] interval; that is, $0 \leq I_i^k \leq 1 \;\forall\, i = 1, \ldots, N$. In this case one can verify that $E^k = \frac{I^k}{S^k}$ where $S^k = \sum_{i=1}^{N} s_i^k$ and $u_i^k = \frac{s_i^k}{S^k}$. This implies that the K&H model delivers values of the composite indicator that are close but not always equal to those obtained from the BoD model. The two models result in the same estimate of the composite indicator if the weights in the BoD model sum up to one as it is by default the case in the K&H model. Notice though that $\sum_{i=1}^{N} s_i^k \geq 1$ when $0 \leq I_i^k \leq 1$.

More importantly, once we have estimated the weights in the BoD model we can use them to obtain the weights of the K&H model but the opposite is not possible. One can verify that the former is straightforward by using the relation $u_i^k = \frac{s_i^k}{S^k}$ $\forall\, k = 1, \ldots, K$ while it is clear that we cannot obtain the s_i^k's from the u_i^k's because

$s_i^k = u_i^k S^k$ consists a system of linearly dependent equations since $\sum_{i=1}^{N} u_i^k = 1$. This asymmetry is due to the type and the number of constraints in the K&H model. For the same reason, however, the K&H model is computationally less demanding. The bottom line though is that anytime we estimate the BoD model with all sub-indicators normalized to lie in the [0,1] interval we can also impute the E^k's while if we estimate the K&H model we cannot impute the I^k's.

Another issue is related to the aggregation (across DMUs) properties of the two models. For the K&H model these are not yet known and for this reason we are unable to integrate the analysis of research productivity from the individual to the institution (department or university) level in a theoretically consistent way. In contrast, such an aggregation rule for the BoD model is developed by Karagiannis (2016) within the framework of aggregate efficiency scores. There it is shown that the arithmetic average is the theoretically consistent aggregation rule for the BoD model;[8] that is:

Thus, the aggregate composite performance indicator equals the simple (un-weighted) arithmetic average of the estimated composite indicators across all DMUs included in the sample. This result stems from the denominator rule and the fact that the BoD model is essentially a radial, input-oriented model with a single constant input. The denominator rule states that consistency in aggregation across DMUs of ratio-type performance measures, such as efficiency or effectiveness indices, is ensured as long as the aggregation shares are defined in terms of the variable being in the denominator (Färe and Karagianis, 2017). For an input-oriented model as the BoD, where efficiency is defined by the ratio of potential to actual input, the aggregation shares are in terms of actual cost or input shares. But since all evaluated DMUs have the same amount of the single input, namely one unit, and assuming that they face the same price for it, the input share of each DMU is equal to 1/K. In terms of research performance, eq. (4) implies that department's research productivity can be simply estimated by means of the average research productivity of its faculty members.[9]

2.2 *Common Weights*

Estimation of research effectiveness by using common instead of variable weights implies special features for both the BoD and the K&H models. First, for the BoD model Karagianis and Paleologou (2014) shown that common weights are related to average cross efficiency. In particular, they show that average cross efficiency in the BoD model is given as:

$$\bar{I}^k = \frac{1}{K}\sum_{j=1}^{K} I_j^k = \frac{1}{K}\sum_{j=1}^{K}\sum_{i=1}^{N} s_i^j I_i^k = \sum_{i=1}^{N} \bar{s}_i I_i^k \qquad (5)$$

where $I_j^k = \sum_{i=1}^{N} s_i^j I_i^k$ is k[th]'s DMU cross efficiency defined in terms of the j[th]'s DMU weights. That is, average cross efficiency in the BoD model is based on a set of common weights given by the simple arithmetic average of the weights obtained from the self-appraisal version of the model as given in (1). Notice that the presence of constant returns of scale eliminates the possibility of negative cross efficiencies in an input-oriented model such as the BoD.[10] However, it is possible that the efficient DMUs do not have a unique set of optimal weights. According to Lins *et al.* (2003) this may not be a serious problem when the number of efficient DMUs is small and thus their influence in the calculation of the average weights is small. Otherwise, a secondary goal scheme (see e.g. Doyle and Green (1994b), Liang *et al.* (2008) and Wang and Chin (2010)) has to be used to guarantee uniqueness of the estimated output weights. Doyle and Green (1994b) aggressive and benevolent formulations are among the most widely used ones. Lastly, notice that it is possible that none of the effectiveness scores calculated from eq. (5) is equal to one and thus in the case of common weights there may not be an efficient DMU.

On the other hand, a set of common weights for the K&H model can be obtained by applying the compromise solution approach proposed by Kao and Hung (2005). In particular, we may estimate econometrically a linear ordinary least squared (OLS) model (not including an intercept term) with the composite indicator obtained from the conventional form of the model as given in (3) to be the dependent variable and the set of sub-indicators to be the independent variables, under the restriction that the estimated parameters sum up to one. That is,

$$E^k = \sum_{i=1}^{N} \bar{u}_i I_i^k + \varepsilon^k \qquad (6)$$

where $\sum_{i=1}^{N} \bar{u}_i = 1$ and $\epsilon^k \sim N(0, \sigma^2)$. Then, the composite indicator of the k[th] DMU is obtained as $\tilde{E}^k = \sum \bar{u}_i I_i^k$.

2.3 *Restricted Weights*

The last set of weights on which a research assessment may be based on is that of reflecting value judgment. For the BoD model, this involves weight restrictions in the multiplier form of the model. Several types of weight restrictions have been proposed for this purpose including pie shares (Cherchye *et al.*, 2007) and partial descending ordering restrictions. In the latter, the importance of sub-indicators is ranked in a sequence rather than by specifying exact weights or a range of weight values. This weak partial ordering of weights has the advantage of being simple but it does not provide information on how much more important one sub-indicator is in relation to all others (Joro and Viitala, 2004). Nevertheless, it seems a suitable and useful value judgment mechanism in the context of assessing research productivity as it is more convenient for faculty members, internal management (e.g., department's head, school's dean, etc.) and other interest groups (e.g., external evaluation committees, stakeholders, etc.) to rank the relative importance of different research outputs rather than to specify their value share either exactly or within a certain range. The partial descending order weights are incorporated into the BoD model by augmenting (1) with the restriction $s_1^k \geq s_2^k \geq s_3^k \geq \cdots$.

On the other hand, the use of partial descending order weights restrictions in the K&H model simplifies to a great extent the estimation of (3). According to Ng (2007, 2008) in this case there is no need to estimate (3) by means of a linear programming problem but rather to impute E^k based on partial averages. In particular, the composite indicator of the k[th] DMU is given as:

$$E^k = \max\left\{I_1^k, \left(\sum_{i=1}^{2} I_i^k\right)/2, \left(\sum_{i=1}^{3} I_i^k\right)/3, \ldots\right\} \qquad (7)$$

While easy to implement, Ng (2007, 2008) approach does not provide estimates of the underlying weights u_i^k. These can be obtained only if the conventional K&H model as given in (3) is augmented with the restriction $s_1^k > s_2^k > s_3^k > \cdots$.

The three variants of the BoD and the K&H models with respectively variable, common and restricted weights are summarized in Table 1. All specifications have the same data requirements, namely a number of sub-indicators referring in our case to different research outputs, and all but the common- and the restricted-weights K&H models are estimated using linear programming. Besides these similarities, the six alternative specifications reflect different assessment paradigms. In particular, the conventional forms of the BoD and the K&H models provide to each faculty member the exclusive right to weight his/her research outputs to his/her best interest (i.e., to maximize overall performance), even if this implies that the weights attached to some output(s) will be zero. On the other hand, the common-weights specification of the BoD and the K&H models evaluate and rank all faculty members using the same "value system" as it is reflected on the average of the estimated output weights. Lastly, the restricted-weights specification of the BoD and the K&H models involve some value judgment in terms of the relative importance of different research outputs that in turns reflects the research priorities of interest group(s), one of which may even be that of the evaluated DMUs.

3. Data Description and Definition of Model Variables

We employ the aforementioned six alternative specifications to evaluate research achievements of the faculty members in the Department of Economics at the University of Macedonia, Greece during the period 2000-2006. In the proposed setting, we consider two research outputs, namely, journal articles and all other publications. As "journal articles" we count all publications in outlets referenced in the Journal of Economic Literature and as "other publications" are considered papers published in journals not referenced in the Journal of Economic Literature and chapters in books and conference volumes. We completely disregard books (textbooks and others) because in Greece almost all of them have not gone through a formal referee process. All data are taken from the *Guide to Published Research Work* of the University of Macedonia (2009).

We construct the two research sub-indicators, for 'journal articles' and 'other publications', by using the whole counting method, where each co-author receives full credit.[11] Thus, a paper co-authored by two or more faculty members of the same department will appear in the publication record of all of them. As a result, the whole counting method yields the largest possible volume of research output and in evaluations at the faculty level, tends to favor those who collaborate with others (Hagen, 2014). This practice is however quite common to the economics profession. Hence our assessment provides, in a sense, the most optimistic evaluation of faculty members in question. To ensure that the values of the two research sub-indicators lie in the [0,1] interval, we normalize the values of the research outputs obtained from the whole counting method by using the 'distance to the group leader' formula, namely $y_i^k / \max_k y_i^k$ for $i=1,2$, where y refers to 'journal articles' and 'other publications'. That is, the elements of the resulting 'journal articles' and 'other publications' vectors (with entries corresponding to faculty members) are divided respectively by their maximum value in the sample.

The number of faculty members ranged from 22 in 2000 to 25 in 2006, with three of them joining the department after 2003. There were two faculty members that had no journal article publication during the period 2000-2006 and only one publication in the category of "other publications". For all but one of the sample years, these faculty members violate the minimum data requirement, namely that of having at least one positive output per year and hence they were excluded to avoid problems in estimating the models in (1) or (3). In addition, we disregard from the evaluation the three faculty members that jointed the department after 2003 and for whom there were no data for the whole period under consideration.

Even though there are no *a priori* norms about the length of the assessment period, empirical evidence from bibliometric studies (see e.g. Abramo, D'Angelo and Cicero, 2012) and practical reasons suggest that usually to be 3 to 5 years, depending on the scientific discipline considered. The practical reasons are related to high volatility of publications between adjacent years, which may be due to both random factors as well as the different procedures (e.g., number of referees, review rounds, etc.) used by the journals that are going to affect the time required from a paper's date of submission to its acceptance and then from acceptance to the actual publication date. The impact of these differences is greater when annual data are used and/or the

assessment is at the individual (i.e., faculty member) level since they tend to increase the number of zero entries in the research output vectors. To deal with this problem, we use three-year moving averages running from 2000-02 to 2004-06, which makes the assessment period triennial.

Summary statistics for the data are given in Table 2. The three-year moving average of 'journal articles' in the first two assessment periods is around one per faculty member while it increases steadily to one and two thirds in the last assessment period. The maximum is four 'journal articles' in the first two assessment periods and it increases only slightly to four and two thirds in the last assessment period. In contrast, the minimum value ranges from zero in the first two assessment periods (as respectively 10% and 15% of faculty members were un-productive) to one third in the last three periods. On the other hand, the three-year moving average of the category of 'other publications' declines from 0.717 in the first assessment period to 0.533 in the last assessment period. Its maximum value increases from three publications in the first assessment period to four in the third period and then decreases again to three in the last assessment period. However, in all assessment periods, there are a quite large number of faculty members (ranging from 25% to 35%) that have no 'other publications'. Also notice that for all assessment periods and both research outputs, the three-year moving average is greater than its median, indicating a rather skewed distribution. Moreover, for all assessment periods, the dispersion (measured by the standard deviation) of 'journal articles' is greater than that of 'other publications'.

4. Empirical Results

Frequency distributions and summary statistics of the research effectiveness scores at the faculty level are given in Tables 3 and 4. In Tables 3 we report the estimates from the three alternative weighting schemes, identified as variable (I), common (II) and restricted (III), for the BoD model and in Table 4 are the corresponding figures from the K&H model. In the computation of average cross efficiencies for the common-weights BoD model we have used the output multipliers from the conventional model as the number of efficient faculty members is very small, ranging from one to three depending on the assessment period, and thus their multipliers' are not expected to have a significant impact on average values. On the other hand, in the restricted-weights formulation of both models we have assumed that the importance of 'journal articles' is greater than that of 'other publications'.

From Tables 3 and 4 we can see that irrespectively of the estimated model and/or the weighing scheme, the research performance of the faculty members is relatively poor and the majority of the effectiveness scores are in the range of 0.1 to 0.5. The distribution of effectiveness scores for all model specifications is found to be highly skewed to the left, with long and thin right tails reflecting the small number of high performing faculty members. Indeed, there is a limited number of faculty members that are found to be efficient, the number of whom in the conventional form of both models is usually two. As it is expected, the number of efficient faculty members decreases as we move from the variable-weights to common- and to restricted-weights scheme, and this is true for both models. It is known that both weight restrictions and cross efficiency have been used in DEA literature for this purpose; namely, to increase the discrimination power of the model and thus to result in a smaller number of efficient DMUs. Notice that the common-weights scheme of both models does not deem efficient any faculty member for three assessment periods (2002-04, 2003-05 and 2004-06).

The simple and rank correlation coefficients, reported in Table 5, show that the six alternative specifications result on average in quite similar effectiveness scores and ranking of faculty members in terms of their research achievements. In most of the cases, correlation coefficients are in the range of 0.9 and above. The great similarity of the effectiveness scores between the BoD and the K&H models is due to the fact that the sum of the estimated weights in the BoD model is in many cases equal to one, as it is by default the case in the K&H model. On the other hand, the changes in ranking among the alternative models are mainly in the range of one to two positions, with the exceptions of few faculty members who may change up to six positions. The research portfolio of these faculty members is in favor of the 'journal articles' and they have only a limited number of 'other publications'.

The effectiveness scores at the department level and by professional rank are given in Tables 6 and 7. These scores have been computed by using eq. (4). Across models and assessment periods, the department effectiveness scores are well below 0.5, with a low of 0.255 for the conventional K&H model in the first assessment period and a high of 0.428 for the conventional and the restricted-weights BoD models in the 2003-05 period. Thus, during the sample period, the department has on average achieved far less than half of its potential outputs with its current composition of faculty members. In other words, the current faculty members could have

produced on average far more than twice as much research outputs as they actually did. This by almost any standards is considered to be a rather poor performance that in addition, indicates a relatively high degree of heterogeneity among faculty members' achievements. Specifically, there are very few efficient units, a relatively large number of effectiveness scores lower than 0.4, and very few faculty members with scores in the range of 0.7 to 0.9 (see also Table 3 and 4).

The empirical results by professional rank (see Table 7) indicate that there is no uniform ranking pattern. For the first three assessment periods, full professors rank first while in the last two assessment periods, lecturers and associate professors take lead. This is true for all model specifications except for the common-weights K&H model and the restricted-weights BoD and K&H models for the second assessment period. As a result, we may argue that during the sample period there is a rather weak positive association between tenure and research effectiveness but no relationship is found between professional rank and research effectiveness. In addition, in most of the model specifications, either the leading or the first two best performing professional ranks are above the departmental average and the rest are below (see Tables 6 and 7).

The comparison of the alternative models and/or weighting schemes reveals three patterns: *first*, at the department level, effectiveness scores from the BoD model are greater than or equal to those from the K&H model for the variable- and the restricted-weights schemes while there is no clear pattern between the BoD and the K&H effectiveness scores from the common-weights scheme. Thus for the data at hand the BoD model tends to provide a more optimistic evaluation at least for the cases of the variable and the restricted weights. The same seems to be true for the case of common weights for all but the second assessment period when the K&H model delivered higher departmental effectiveness scores.

Second, for both the BoD and the K&H models, the estimated scores with variable weights are greater than or equal to those obtained with restricted weights and those in turn are greater than the effectiveness scores obtained with common weights. In other words, as the values assigned to research outputs are gradually restricted (from being variable, variable within certain limits, and common to all DMUs) the estimated effectiveness scores tend to decrease. The common-weights scheme tends to result in the most pessimistic estimates of research effectiveness. In terms of managerial implications, these imply that research performance appear better

if reasonable weights on research outputs' relative importance are predetermined by internal management (e.g., department's head, school's dean), the government, stakeholders, external evaluation committees, or even the faculty members than if all faculty members have a 'say' in the performance and research priorities of other faculty members.

Third, for both the BoD (with the exception of the last assessment period) and the K&H models, the dispersion of the estimated effectiveness scores from the variable-weights scheme is less than the dispersion of the estimated effectiveness scores from the restricted-weights scheme and that in turn is less than the dispersion of the estimated effectiveness scores from the common-weights scheme. In other words, as the values assigned to research outputs are gradually restricted (from being variable, variable within certain limits, and common to all DMUs) the standard deviation of the estimated scores tends to increase (see Table 3 and 4). The common-weights scheme results in the more dispersed estimates of research effectiveness.

Fourth, it seems that there is more variability in estimated effectiveness scores within than between the two models. That is, the weighting schemes tend to induce more variability in the estimated effectiveness scores than the two models do. This may be expected in our case as the sum of the estimated weights in the BoD model are equal to one for many (but not all) faculty members, as it is by default the case in the K&H model. As a result, the estimated effectiveness scores for each type of *a posteriori* weights were more similar than the effectiveness scores for each model under the three alternative *a posteriori* weighting schemes.

Fifth, all but the common-weights BoD model result in the same temporal pattern of research effectiveness. Based on departmental scores, we can infer that research effectiveness first increases and then decreases, and this temporal pattern is repeated again in the following two assessment periods (see Table 6).

Sixth, the estimated output multipliers from the variable- and the common-weights formulations of both the BoD and the K&H models confirm in general terms the relative importance of research output assigned by the assumed partial descending ordering, namely that the relative importance of 'journal articles' is greater than that of 'other publications' (see Table 8). In addition, we can see that the relative importance of 'journal articles' tend to increase over time and that of 'other publications' to decrease. The largest gap in the relative importance of the two research outputs appear in the restricted-weights K&H model for the fourth

assessment period and the smallest one in the common-weights K&H model for the first assessment period.

5. Concluding Remarks

In this paper we examine three alternative *a posteriori* weighting schemes with variable, common and restricted weights schemes in order to assess research performance by means of two seemingly similar non-parametric models, namely the BoD and the K&H model. For the BoD model it turns out that the three alternative *a posteriori* weighting schemes corresponds to three different aspects of performance evaluation based respectively on self-appraisal, peer-appraisal and value judgment. In the case of the K&H model, the variable-, the common- and the restricted-weights scheme corresponds respectively to the conventional formulation of the model, the compromise solution approach proposed by Kao and Hung (2005), and the partial average formulation suggested by Ng (2007, 2008).

Our empirical results, based on the publication records of the faculty members in the department of Economics at the University of Macedonia, Greece during the period 2000-2006, show that there is more variability in the estimated effectiveness scores among the alternative weighting schemes within each model rather than between models for any particular weighting scheme. In particular, for both the BoD and the K&H models, the estimated scores from the variable-weights scheme are greater than or equal to those obtained from the restricted-weights scheme and those in turn are greater than the effectiveness scores from the common-weights scheme. On the other hand, we also found that the effectiveness scores from the BoD model are greater than or equal to those from the K&H model for the variable- and the restricted-weights schemes while there is no clear pattern between the BoD and the K&H effectiveness scores from the common-weights scheme.

Table 1: Models Representation

	Variable Weights	Common Weights	Restricted Weights
BoD	conventional model as in eq. (1)	average cross efficiency as in eq. (5)	eq. (1) augmented with the restrictions $s_1^k > s_2^k > s_3^k > \cdots$
K&H	conventional model as in eq. (3)	compromise solution approach as in eq. (6)	partial average as in eq. (7)

Table 2: Descriptive Statistics of Model Variables

	2000-02	2001-03	2002-04	2003-05	2004-06
			Journal Articles		
Average	1.083	1.050	1.350	1.433	1.667
Median	0.667	1.000	1.333	1.000	1.333
St.dev	1.037	0.818	1.029	1.021	1.170
Max	4.000	3.000	4.333	3.667	4.667
Min	0.000	0.000	0.333	0.333	0.333
% Zeros	10	15	0	0	0
			Other Publications		
Average	0.717	0.600	0.633	0.600	0.533
Median	0.333	0.333	0.333	0.333	0.333
St.dev	0.926	0.754	0.923	0.835	0.712
Max	3.000	2.667	4.000	3.667	3.000
Min	0.000	0.000	0.000	0.000	0.000
% Zeros	30	30	25	30	35

Table 3: Frequency Distribution of Faculty Research Effectiveness based on the BoD Model, Department of Economics, University of Macedonia, Greece, 2000-2006.

		2000-02	2001-03	2002-04	2003-05	2004-06
BoD I	(0,0.1)	1	0	2	0	2
	[0.1,0.2)	6	3	3	4	2
	[0.2,0.3)	4	5	3	4	4
	[0.3,0.4)	2	4	6	6	5
	[0.4-0.5)	0	2	2	1	2
	[0.5,0.6)	4	2	1	1	0
	[0.6,0.7)	1	1	1	0	2
	[0.7,0.8)	0	0	0	0	0
	[0.8,0.9)	0	2	0	0	1
	[0.9,1)	0	0	0	1	0
	1	2	1	2	3	2
	Max	1.000	1.000	1.000	1.000	1.000
	Min	0.080	0.130	0.080	0.120	0.070
	Median	0.235	0.330	0.310	0.305	0.352
	Stdev	0.280	0.263	0.269	0.300	0.277
BoD II	(0,0.1)	5	2	5	0	3
	[0.1,0.2)	6	7	1	4	3
	[0.2,0.3)	3	2	8	7	5
	[0.3,0.4)	1	4	1	3	4
	[0.4-0.5)	2	3	2	2	0
	[0.5,0.6)	2	1	1	0	1
	[0.6,0.7)	0	0	1	0	1
	[0.7,0.8)	0	0	0	2	2
	[0.8,0.9)	0	0	0	1	0
	[0.9,1)	0	0	1	1	1
	1	1	1	0	0	0
	Max	1.000	1.000	0.950	0.974	0.934
	Min	0.044	0.056	0.062	0.104	0.061
	Median	0.194	0.240	0.247	0.284	0.271
	Stdev	0.232	0.228	0.224	0.263	0.246
BoD III	(0,0.1)	4	2	2	0	2
	[0.1,0.2)	6	3	3	4	2
	[0.2,0.3)	1	3	3	4	5
	[0.3,0.4)	3	4	6	6	4
	[0.4-0.5)	0	3	2	1	2
	[0.5,0.6)	4	1	1	1	0
	[0.6,0.7)	1	2	1	0	2
	[0.7,0.8)	0	0	0	0	0
	[0.8,0.9)	0	1	0	0	1
	[0.9,1)	0	0	1	1	0
	1	1	1	1	3	2
	Max	1.000	1.000	1.000	1.000	1.000
	Min	0.055	0.065	0.080	0.120	0.070
	Median	0.250	0.330	0.310	0.305	0.348
	Stdev	0.249	0.260	0.261	0.300	0.279

Table 4: Frequency Distribution of Faculty Research Effectiveness based on the K&H Model, Department of Economics, University of Macedonia, Greece, 2000-2006.

		2000-02	2001-03	2002-04	2003-05	2004-06
K&H I	(0,0.1)	1	0	5	2	2
	[0.1,0.2)	6	3	0	3	2
	[0.2,0.3)	4	5	4	7	5
	[0.3,0.4)	2	4	5	2	4
	[0.4-0.5)	0	2	2	1	2
	[0.5,0.6)	4	2	1	1	0
	[0.6,0.7)	1	1	1	0	2
	[0.7,0.8)	0	0	0	0	1
	[0.8,0.9)	0	2	0	0	0
	[0.9,1)	0	0	0	1	0
	1	2	1	2	3	2
	max	1.000	1.000	1.000	1.000	1.000
	min	0.110	0.130	0.080	0.090	0.070
	median	0.250	0.330	0.310	0.270	0.330
	Stdev	0.280	0.263	0.275	0.310	0.271
K&H II	(0,0.1)	6	3	5	2	3
	[0.1,0.2)	6	5	2	3	3
	[0.2,0.3)	2	3	7	7	5
	[0.3,0.4)	2	3	2	2	4
	[0.4-0.5)	1	2	1	1	0
	[0.5,0.6)	2	2	2	2	2
	[0.6,0.7)	0	0	0	0	2
	[0.7,0.8)	0	1	0	0	0
	[0.8,0.9)	0	0	1	1	0
	[0.9,1)	0	0	0	2	1
	1	1	1	0	0	0
	max	1.000	1.000	0.888	0.958	0.907
	min	0.056	0.023	0.060	0.091	0.061
	median	0.179	0.294	0.239	0.266	0.261
	Stdev	0.236	0.245	0.208	0.263	0.232
K&H III	(0,0.1)	4	2	5	2	3
	[0.1,0.2)	6	3	1	3	1
	[0.2,0.3)	1	3	3	7	7
	[0.3,0.4)	3	4	5	2	2
	[0.4-0.5)	0	3	2	1	2
	[0.5,0.6)	4	1	1	1	0
	[0.6,0.7)	1	2	2	0	2
	[0.7,0.8)	0	0	0	1	2
	[0.8,0.9)	0	1	0	0	0
	[0.9,1)	0	0	0	1	0
	1	1	1	1	2	1
	max	1.000	1.000	1.000	1.000	1.000
	min	0.056	0.063	0.077	0.091	0.071
	median	0.250	0.333	0.308	0.273	0.286
	Stdev	0.249	0.260	0.247	0.289	0.256

Table 5: Correlation Analysis Results.

		BoD I	BoD II	BoD III	K&H I	K&H II	K&H III
2000-02	BoD I		0.951	0.958	0.998	0.962	0.956
	BoD II	0.900		0.972	0.952	0.960	0.973
	BoD III	0.929	0.959		0.955	0.921	1.000
	K&H I	1.000	0.900	0.929		0.958	0.953
	K&H II	0.914	0.989	0.924	0.914		0.920
	K&H III	0.928	0.960	1.000	0.928	0.925	
2001-03	BoD I		0.967	0.974	0.999	0.957	0.974
	BoD II	0.922		0.972	0.964	0.967	0.972
	BoD III	0.968	0.914		0.973	0.993	1.000
	K&H I	1.000	0.922	0.968		0.954	0.973
	K&H II	0.927	0.941	0.984	0.927		0.993
	K&H III	0.967	0.914	1.000	0.967	0.984	
2002-04	BoD I		0.984	0.998	0.961	0.967	0.972
	BoD II	0.960		0.985	0.964	0.988	0.978
	BoD III	0.998	0.975		0.958	0.969	0.975
	K&H I	0.997	0.958	0.996		0.976	0.985
	K&H II	0.947	0.999	0.965	0.946		0.992
	K&H III	0.966	0.994	0.980	0.970	0.994	
2003-05	BoD I		0.992	0.999	0.960	0.994	0.976
	BoD II	0.991		0.987	0.950	0.997	0.971
	BoD III	1.000	0.991		0.959	0.989	0.971
	K&H I	0.997	0.987	0.997		0.960	0.987
	K&H II	0.960	0.980	0.960	0.961		0.979
	K&H III	0.984	0.990	0.984	0.988	0.992	
2004-06	BoD I		0.980	0.991	0.985	0.960	0.962
	BoD II	0.981		0.995	0.959	0.992	0.988
	BoD III	0.998	0.987		0.971	0.982	0.983
	K&H I	0.995	0.975	0.993		0.938	0.954
	K&H II	0.960	0.995	0.968	0.955		0.993
	K&H III	0.979	0.996	0.983	0.980	0.994	

Note: The simple correlation coefficients are placed below main diagonal and the Spearman rank correlation coefficients above main diagonal.

Table 6: Department Research Effectiveness Scores, 2000-2006.

	variable weights		common weights		restricted weights	
	BoD I	K&H I	BoD II	K&H II	BoD III	K&H III
2000-2002	0.365	0.365	0.258	0.255	0.318	0.318
2001-2003	0.415	0.415	0.294	0.327	0.382	0.382
2002-2004	0.364	0.353	0.297	0.277	0.357	0.332
2003-2005	0.428	0.412	0.376	0.373	0.428	0.402
2004-2006	0.410	0.400	0.349	0.332	0.406	0.378

Table 7: Research Effectiveness Scores by Academic Rank, Department of Economics, University of Macedonia, Greece, 2000-2006.

	Average Efficiency	2000-02	2001-03	2002-04	2003-05	2004-06
BoD I	Professor	0.454	0.478	0.427	0.408	0.322
	Associate Professor	0.334	0.371	0.324	0.444	0.484
	Assistant Professor	0.250	0.440	0.310	0.305	0.430
	Lecturer	0.080	0.330	0.310	0.550	0.360
BoD II	Professor	0.317	0.334	0.331	0.351	0.262
	Associate Professor	0.238	0.272	0.276	0.398	0.418
	Assistant Professor	0.194	0.301	0.272	0.284	0.392
	Lecturer	0.050	0.183	0.247	0.452	0.303
BoD III	Professor	0.399	0.439	0.415	0.408	0.322
	Associate Professor	0.284	0.335	0.320	0.444	0.476
	Assistant Professor	0.250	0.440	0.310	0.305	0.430
	Lecturer	0.080	0.330	0.310	0.550	0.360
K&H I	Professor	0.454	0.478	0.416	0.395	0.319
	Associate Professor	0.334	0.371	0.311	0.426	0.465
	Assistant Professor	0.250	0.440	0.310	0.270	0.430
	Lecturer	0.080	0.330	0.310	0.550	0.360
K&H II	Professor	0.310	0.374	0.308	0.331	0.242
	Associate Professor	0.239	0.288	0.259	0.405	0.401
	Assistant Professor	0.179	0.384	0.258	0.266	0.384
	Lecturer	0.041	0.271	0.239	0.503	0.307
K&H III	Professor	0.399	0.440	0.376	0.369	0.284
	Associate Professor	0.283	0.335	0.302	0.427	0.451
	Assistant Professor	0.250	0.444	0.308	0.273	0.429
	Lecturer	0.083	0.333	0.308	0.545	0.357

Table 8: Average output weights per Model

	2000-02	2001-03	2002-04	2003-05	2004-06
			Journal Articles		
BoD I	0.600	0.550	0.804	0.829	0.848
BoD II	0.600	0.550	0.804	0.829	0.848
BoD III	0.800	0.775	0.883	0.861	0.911
K&H I	0.600	0.600	0.850	0.950	0.800
K&H II	0.492	0.812	0.776	0.922	0.861
K&H III	0.800	0.800	0.850	0.975	0.900
			Other Publications		
BoD I	0.400	0.450	0.291	0.319	0.259
BoD II	0.400	0.450	0.291	0.319	0.259
BoD III	0.200	0.225	0.233	0.301	0.223
K&H I	0.400	0.400	0.150	0.050	0.200
K&H II	0.508	0.188	0.224	0.078	0.139
K&H III	0.200	0.200	0.150	0.025	0.100

References

Abramo, G., D'Angelo, C.A. and T. Cicero. What is the appropriate length of the publication period over which to assess research performance?, *Scientometrics*, 2012, 93(3), 1005-17.

Abramo, G., D'Angelo, C.A. and F. Pugini. The measurement of Italian universities' research productivity by a non parametric-bibliometric methodology, *Scientometrics*, 2008, 76(2), 225-44.

Abramo, G., Cicero, T. and C.A. D'Angelo. A field-standardized application of DEA to national-scale research assessment of universities, *Journal of Informetrics*, 2011, 5(4), 618-28.

Caporaletti, L.E., Dula, J.H. and N.K. Womer. Performance evaluation based on multiple attributes with nonparametric frontiers, *Omega*, 1999, 27(6), 637-45.

Charnes, A., Cooper, W.W. and E. Rhodes. Measuring the efficiency of decision making units, *European Journal of Operational Research*, 1978, 2(6), 429-44.

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

de Witte, K. and L. Hudrlikova. What about excellence in teaching? A benevolent ranking of universities, *Scientometrics*, 2013, 96(1), 337-64.

de Witte, K. and N. Rogge. To publish or not to publish? On the aggregation and drivers of research performance, *Scientometrics*, 2010, 85(3), 657-80.

Doyle, J.R. Multiattribute choice for the lazy decision maker: Let the alternatives decide! *Organizational Behavior and Human Decision Processes,* 1995, 62(1), 87-100.

Doyle, J.R. and R. Green. Self and peer appraisal in higher education, *Higher Education*, 1994a, 28(2), 241-64.

Doyle, J.R. and R. Green. Efficiency and cross efficiency in DEA: Derivation, meanings and uses, *Journal of Operational Research Society*, 1994b, 45(5), 567-78.

Färe, R. and G. Karagiannis. The denominator rule for share-weighting aggregation, *European Journal of Operational Research*, 2017 (forthcoming).

Hagen, N.T. Counting and comparing publication output with and without equalizing and inflationary bias, *Journal of Informetrics*, 2014, 8(2), 310-17.

Joro, T. and E.J. Viitala. Weight-restricted DEA in action: From expert opinions to mathematical models, *Journal of Operational Research Society*, 2004, 55(8), 814-21.

Karagiannis, G. On aggregate composite indicators, *Journal of Operational Research Society*, 2017 (forthcoming).

Karagiannis, G. and C.A.K. Lovell. Productivity measurement in radial DEA models with a single constant input. *European Journal of Operational Research*, 2016, 251(1), 321-28.

Karagiannis, G. and S.M. Paleologou. Towards a composite public sector performance indicator, paper presented in the *2014 Asia Pacific Productivity Conference*, Brisbane, July 2-4, 2014.

Kao, C. Evaluation of junior colleges of technology: The Taiwan case, *European Journal of Operational Research*, 1994, 72(1), 43-51.

Kao, C. and H.T. Hung. Ranking University Libraries with a posteriori Weights, *Libri*, 2003, 53(4), 282-89.

Kao, C. and H.T. Hung. Data envelopment analysis with common weights: The compromise solution approach. *Journal of Operational Research Society*, 2005, 56(10), 1196-1203.

Kao, C. and H.Y Hung. Management performance: An empirical study of the manufacturing companies in Taiwan, *Omega*, 2007, 35(2), 152-160.

Kao, C., Liu, S.T. and H.L. Pao. Assessing improvement in management research in Taiwan, *Scientometrics*, 2012, 92(1), 75-87.

Kao, C. and H.L. Pao. An evaluation of research performance in management of 168 Taiwan universities, *Scientometrics*, 2009, 78(2), 261-77.

Kao, C., Wu, W.Y., Hsieh, W.J., Wang, T.Y., Lin, C. and L.H. Chen. Measuring the national competitiveness of Southeast Asian countries, *European Journal of Operational Research*, 2008, 187(2), 613-28.

Liang, L., Wu, J., Cook, W.D. and J. Zhu. Alternative secondary goals in DEA cross-efficiency evaluation, *International Journal of Production Economics*, 2008, 113(2), 1025-1030.

Lim, S. and J. Zhu. DEA cross-efficiency evaluation under variable returns to scale. *Journal of Operational Research Society*, 2015, 66(3), 476-87.

Lins M.P.E., Gomes, E.G., Soares de Mello, J.C.C.B. and A.J.R. Soares de Mello. Olympic ranking based on a zero sum gains DEA model. *European Journal of Operational Research*, 2003, 148(2), 312-32

Liu, W.B., Zhang, D.Q., Meng, W., Li, X.X. and F. Xu. A study of DEA models without explicit inputs, *Omega*, 2011, 39(5), 472-80.

Lovell, C.A.K. and J.T. Pastor. Radial DEA models without inputs or without outputs, *European Journal of Operational Research*, 1999, 118(1), 46-51.

Murias, P., de Miguel, J.C. and D. Rodriguez. A composite indicator for university quality assessment: The case of Spanish Higher Education System, *Social Indicator Research*, 2008, 89(1), 129-46.

Ng, W.L. A simple classifier for multiple criteria ABC analysis, *European Journal of Operational Research*, 2007, 177(1), 344-353.

Ng, W.L. An efficient and simple model for multiple criteria supplier selection problem, *European Journal of Operational Research*, 2008, 186(3), 1059-67.

Oral, M., Oukil, A., Malouin, J.L. and O. Kettani. The appreciative democratic voice of DEAL The case of faculty academic performance evaluation, *Socio-Economic Planning Sciences*, 2014, 48(1), 20-28.

Prieto, A.M. and J.L. Zofio. Evaluating effectiveness in public provision of infrastructure and equipment: The case of Spanish municipalities, *Journal of Productivity Analysis*, 2001, 15(1), 41-58.

Soares de Mello, J.C.C.B., Angulo Meza, L.A., Quintanilha da Silveira, J. and E. Concales Gomes. About negative efficiencies in cross efficiency BCC input oriented models. *European Journal of Operational Research*, 2013, 229(3), 732-37.

Thanassoulis, E., de Witte, K., Johnes, G., Johnes, J., Karagiannis, G. and M.C.S. Portela. Applications of Data Envelopment Analysis in education, in Zhu, J. (ed), *Data Envelopment Analysis: A handbook of empirical studies and applications*, Springer, 2016, 367-438.

Wang, Y.M. and K.S. Chin. Some alternative models for DEA cross-efficiency evaluation, *International Journal of Production Economics*, 2010, 128(1), 332-38.

Wang, Y.M., Luo, Y. and Y.X. Lin. Common weights for fully ranking decision making units by regression analysis, *Expert Systems with Applications*, 2011, 38(8), 9122-28.

Footnotes

[1] Besides research productivity, the BoD and the K&H models have been used in a number of other applications, including cases where performance evaluation is based on ratio variables or on targets set by management, construction of composite indicators, and several multiple criteria decision making analysis problems such as inventory classification, supplier selection, and service quality; references for such studies can be found in Karagiannis and Lovell (2016). The results of the present study will therefore be of interest to analysts of all these fields.

[2] This is apparently true for any other applications of the BoD and the K&H models. In that sense the present work is to the best of our knowledge the first to compare and contrast these variants of the BoD and the K&H models.

[3] There are also some studies on research efficiency that account for the resources employed in research activities; see Kao (1994), Abramo, D'Angelo and Pugini (2008), Murias, de Miguel and Rodriguez (2008), Abramo, Cicero and D'Angelo (2011), and de Witte and Hudrlikova (2013). There is also a large number of studies at the department or at the university level that do not evaluate research *per se* but consider it as one element of the output set, along with teaching, administration services, etc.; an extensive list of such studies can be found in Thanassoulis *et al.* (2016).

[4] The other two research productivity evaluation methods, namely peer review and bibliometrics, rely respectively on *a priori* weights reflecting experts' or stakeholders' opinions or use equal weights and appropriate normalizations/standardization to obtain comparable metrics.

[5] Another advantage of the common-weights scheme is that it can be applied to assess performance for DMUs not being in the sample (Kao and Hung, 2007).

[6] More on the radial DEA models with a single constant input can be found in Lovell and Pastor (1999), Caporaletti, Dula and Womer (1999) and Liu *et al.* (2011). Notice also that unitary input DEA models are equivalent to DEA models without explicit inputs.

[7] This in turn implies that the composite indicator can be estimated using the output-oriented version of the model. The latter may have a more appealing interpretation in terms of efficiency measurement when there is a single unitary input. However, if

someone is interest in estimating the weights assigned to each sub-indicator then this can only be done with the input-oriented formulation and in particular, its multiplier form.

[8] By consistency we mean that the resulting aggregate measure has exactly the same intuitive interpretation as the individual efficiency scores.

[9] Notice that eq. (4) holds regardless of whether we impose or not weight restrictions when estimating the composite indicator of research productivity.

[10] The input-oriented model with variable returns to scale may generate negative cross efficiencies for the DMUs exhibiting decreasing returns to scale (Soares de Mello *et al.*, 2013; Lim and Zhu, 2015). The output-oriented model with variable or constant returns to scale does not suffer from this problem.

[11] There are other three counting methods for joint papers (Hagen, 2014), namely, (a) straight counting where the most prominent co-author (being either the first author or the corresponding author) receives full credit and the rest receive none; (b) fractional counting where credit is shared either equally by all co-authors (simple fractional measure) or based on some predetermined weights (full fractional measure); and (c) harmonic counting. The comparison of the BoD and the K&H models under the four alternative counting methods goes beyond the purpose of the present paper and it is left for future work.