Assessing Research Productivity in Higher Education Institutions

Giannis Karagiannis

Professor, Department of Economics, University of Macedonia, 156 Egnatia Str. Thessaloniki, 54006 GREECE; karagian@uom.gr

Assessing Research Productivity in Higher Education Institutions

1. Measuring Inputs and Outputs

Research activity may be viewed as a multi-input, multi-output production process with execution time that notably differs across disciplines and even in fields within disciplines. It involves several forms of (a) human (e.g., academic stuffs, PhD students, research assistants), (b) tangible (e.g., scientific instruments, materials) and (c) intangible (e.g., accumulated knowledge, social networks) resources that are combined to produce an output called "new knowledge", which has also tangible (e.g., publications, patents, conference presentations) and intangible (e.g., tacit knowledge, consulting services) features. Besides these, research output has two other aspects that are of special interest in assessment exercises: quality (i.e., research excellence) and value or impact, with the latter being measured by citations counts when academic impact is concerned and judgmentally when impact is in non-academic (e.g., business or government) domain. The decision-making unit behind research activity varies depending on the level of analysis from individual researchers to institutions, such as departments, schools, or even the university as a whole.

All these make research assessment quite a complicated task requiring several assumptions and simplifications to be made at the outset. The first of them concerns with the length of the assessment period considered, which is directly related to the length of the publication period. Both the period from a paper's date of submission to a journal and its acceptance and the period from acceptance to actual publication date differ even within the same discipline. This may be due to among other things the procedures followed by different journals (e.g., number of referees, review rounds, etc.). The likelihood that such factors affect the measures of research performance tends to increase for relatively short assessment periods. This is particularly true for evaluation exercises conducted at the individual researcher level and less for more aggregated levels of analysis, i.e., departments, schools, or universities. Even though there are no *a priori* norms, empirical evidence from bibliometric studies (i.e., Abramo, D'Angelo and Cicero, 2012) suggest that the preferable assessment period is between three to five years, depending upon the scientific discipline considered.

On the input side, measurement of production factors other that labor is in most of the cases difficult or even impossible due to lack of data. We thus usually

assume that resources available to evaluated units are the same at least within the same field and/or within a given institution. These include supporting staff, PhD students and external collaborators, scientific instruments, and materials. In addition, we assume unless data are available that the hours available for research are the same for each individual in a given field category. This is a reasonable assumption for non-competitive higher education systems (e.g., Italy and Greece) where hours devoted to teaching are established by national regulations and are the same for all, regardless of academic ranks. In this case, research can be evaluated separately from teaching as labor is allocated between research and teaching in fixed proportions. It is a less reasonable assumption for competitive higher education systems (e.g., UK, USA and Canada), where there is a trade-off between research and teaching time, and this should explicitly be taken into account in the assessment exercise.

However, the value of time is different and is reflected in the labor cost that varies across academic rank. Since salaries of full, associate and assistant professors differ it would be appropriate to distinguish between them by including three different "types" of labor in the assessment exercise or to measure labor input by its cost if information on individual salaries is available. The purpose of this is to distinguish between different degrees of quality among the employed human resources. Using uniform labor input instead of labor cost as is expected to have a more noticeable impact at the individual researcher rather than at more aggregated (i.e., department, school or university) level.[1] For evaluation studies at the individual researcher level when information on salaries is not available or salaries are equal within academic ranks as in most non-competitive higher education systems (e.g., Italy and Greece), the second best option is to evaluate research productivity by academic rank (Abramo, Cicero and D'Angelo, 2013). Nevertheless, empirical evidence from a bilbiometric study by Abramo, D'Angelo and Solazzi (2010) indicate that the effect of switching from uniform labor input to cost of labor seems to be minimal expect for outliers.

On the output side, as the intangible counterpart of research output is hard to measure we consider only codified new knowledge in assessment exercises. These include articles in academic journals, research monographs, patents awards, and

[1] Since available data show that more senior academic staff have more, better and highly valued (cited) publications, department or university rankings based on uniform labor input will favor units with greater concentration at higher academic ranks.

presentations in conferences as well as their relative importance that differs by subject category and/or its discipline. The most prevalent form of codification for research output is publications in academic journals, which is considered as an acceptable approximation of research output in many fields but less so in the arts, humanities and a good part of social sciences (Abramo, D'Angelo and Di Costa, 2014). But as patents are often followed by publications that describe their content in the relevant scientific area and conference presentations usually precede publication of research work, consideration of the number of publications alone to approximate research output may actually avoid in many cases a potential double counting. Publications may be further distinguished by the type of outlet where are published into academic journals, chapters in edited books or proceedings, research monographs, reports, theses etc.

The way we count publications may induce biases in performance evaluation as the number of co-authors as well as the quality/prestige of the publication outlet are factors that one may want to control for in measuring research output.[2] Following Lin, Huang and Chen (2013) and Hagen (2014) there are four counting methods for collaborative papers: (1) *whole counting* where each collaborating author receives full credit. (2) *Straight counting* where the most prominent collaborator (being either the first author or the corresponding author) receives full credit and the rest receive none. (3) *Fractional counting* where credit is shared either equally by all collaborators (*simple fractional measure*) or based on some predetermined weights (*full fractional measure*). Simple fractional counting is appropriate when authors are listed in alphabetical order or when it is explicitly stated that authorship is equally shared. Full fractional counting may be based on weights provided by field experts or some other authority. (4) *Harmonic counting* where credit is determined as $\frac{\frac{1}{i}}{1+\left(\frac{1}{2}\right)+\left(\frac{1}{3}\right)+\cdots+\left(\frac{1}{N}\right)}$ with i being the position of an author in the byline and N the number of collaborators.[3] For example, for a two-authors paper, the first authors

[2] *A priori* the quality of a publication is independent of the number of collaborators and thus we have to adjust publications counts by both factors.

[3] Hagen (2014) also provided the corresponding formula for harmonic counting in fields like medicine where senior authorship is usually assigned to the first and last collaborator, who are respectively the leader of the specific research and the leader of the entire research group.

receive 2/3 and the second 1/3 of credit. For a paper where three authors involved, the first author receives 6/11, the second 3/11 and the third 2/11 of credit.

One can easily verify that the whole counting method results in the largest possible volume of research output, inducing what is called inflationary bias, and in evaluations at the individual level tends to favor researchers with many co-authors (Hagen, 2014). On the other hand, the simple factional counting method suffers from what is called equalizing bias (i.e., inaccuracy in accreditation when credit is divided equally among co-authors who however have not contributed equally) and in research assessment at the individual level tends to favor secondary authors (Hagen, 2014). For research assessment at the institution (i.e., departments, universities) level, Lin, Huang and Chen (2013) argued that straight and fractional counting are better choices than whole counting. However, Abramo, D' Angelo and Rossati (2013) tend to agree with this only if fractional counting is based on full fractional measures.[4]

In addition, publications counting and accreditation also depend on the level of aggregation that the assessment exercise is conducted. At the department level, research output is equal to the sum of publications with at least one author belonging to this particular department. Notice however that a publication co-authored by researchers of the same department or university is considered only once in research assessment at the department/university level but it accounts for a particular fraction at the individual level. Similarly, a publication co-authored by researchers from different departments of the same university, it will be considered only once in the evaluation of the university but it will account for a particular fraction in assessing performance at the department or individual level. Nevertheless, a publication co-authored by researchers from different universities will account for a particular fraction even for university level evaluations besides its fractional contribution for department or individual level evaluations.

On the other hand, two alternatives have been proposed in the literature to adjust research output for quality: (a) the journal's impact factor and (b) articles citation counts. Even though many will argue in favor of the latter, the reliability of citation counts in representing a quality ladder of an academic article depends on the

[4] There is also a disagreement on whether the choice of counting method affects more papers or citations counts. Lin, Huang and Chen (2013) found that it impacts citation counts more than paper counts while Abramo, D'Angelo and Rosatti (2103) reached the opposite conclusion.

time elapse between the publication date and the moment of observing the number of citations received. Citations observed at a moment too close to the date of publication will not necessarily offer a quality proxy that is preferable to journal impact factor. According to Abramo, D'Angelo and Di Costa (2010), if we do not have data on citations counts for at least a period of two to five years (depending on the scientific field) after the end of the evaluation period considered it will be preferable to use journal impact factor to approximate publication quality. Nevertheless, since the distribution of both citations and journal impact factors are typically skewed to the right in all scientific fields it seems appropriate to use the percentile as means of standardization (Abramo, D'Angelo and Di Costa, 2010a, b).[5]

However if there are data for a sufficient period of time after the end of the evaluation period considered, citations counts can be used not only as a quality ladder to adjust publication counts but also as an addition research output metric accounting for the value of scientific achievements. Academic publications embedding new knowledge have different values measured by their impact on scientific achievements. Citations represent a proxy measure of the value of research output that is usually included in assessment exercises, in spite of limitations due to negative citations and network citations. Notice that when counting citations different weights may be given depending upon the citing article influence, the journal in which it is published, etc.

In order to make citation counts a meaningful metric of research value, their total number should be standardized especially for comparisons across fields to reflect differences in citation intensity as well as the various degrees of covering of each scientific field in the existing citation databases. This will render citation counts comparable across different research fields and time. Different scaling factors have been proposed in the literature: the average, the geometric mean, the median, the z-score, etc. Due to the (right) skewness of citations' distribution it seems preferable to use the median as scaling factor. Empirical evidence from bibliometric studies (i.e., Abramo, Cicero and D'Angelo, 2012a, b) suggest however that the arithmetic average seems the most effective scaling factor when the average is based on the publications actually cited and thus excluding those not cited from the calculation of the arithmetic

[5] Right-hand skewness implies that most papers are relatively little cited and there are only few papers with many citations, and that the vast majority of papers is published in relatively low impact journals.

average.[6] Scaling of citation counts is carried out by multiplying the citations of each publication by the chosen scaling factor that characterizes the distribution of citations of articles from the same scientific field and the same year.

2. Alternative DEA Models for Assessing Research Productivity

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 efficiency score or a composite indicator, reflecting relative achievement.[7] The *a posteriori* weights may be variable (i.e., unit-specific) or common, and may or may not reflect (at least partially) experts' or stakeholders' opinions.

The flexible (unit-specific) weights resulting from conventional DEA models reflect its underlying assumption that each evaluated unit is allowed to choice, under certain regulatory conditions, its own set of input and output weights in order to show in the best possible light relative to other units 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. But if after that is still weak relative to other units in the sample this cannot be put down to the choice of input and output weights.

On the other hand, some authors have argued 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. Several variants/modifications, such as common weights DEA and average cross efficiency, have been used for such purposes (see for example Oral *et al.*, 2014). Lastly, a combination of *a posteriori* and *a priori* weights (i.e., model- and experts/stakeholders-based) may also be possible. Two rather distinct approaches have been used in this respect: peer appraisal by means of cross efficiencies and value judgment DEA. In the former

---

[6] Abramo, Cicero and D'Angelo (2012a, b) also provide empirical evidence indicating that rankings of individual researchers obtained under different scaling factors (i.e., average, median, cited papers average, cited papers median) do not show significant discrepancies.

[7] 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.

case, the value norms (i.e., the DEA weights) of all evaluated units are taken into account when assessing the performance of each unit. In the latter case, 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 the DEA weights accord with intuition.

The second methodological aspect that has to be considered at the outset of the evaluation process is whether resources related to research activities will be taken into account or not. This is equivalent in choosing between efficiency and effectiveness measurement. The former compares the outcome(s) of the research related activities relative to the resources employed for this purpose while the latter compares only the outcome(s) of the research related activities and not the means to achieve them. Conventional DEA models may be used to estimate efficiency of research activities while measuring effectiveness is equivalent to constructing composite performance indicators, which can be done using either DEA-based models such as the benefit-of-the-doubt (BoD) or linear programming models (Kao and Hung, 2003).

The third methodological aspect that has to be considered at the outset of the evaluation process is related to the aggregation level at which the assessment exercise will be conducted. This aggregation level runs from individual level to different degrees of institution/organization aggregation, namely departments, schools/colleges, and the university as a whole. We can thus evaluate research productivity of faculty members as well as of the departments or the universities that belong to. According to Abramo and D'Angelo (2014), for any ranking concerning units that are non-homogenous in their research fields it is necessary to start from the measurement of research productivity at the individual (i.e., faculty members) units and then find an appropriate way to aggregate them. This requires a consistent way to aggregate efficiency and effectiveness scores from the individual to the institution level.

2.1 Measuring Efficiency of Research Activities

Output orientations is appropriate for scientific research since in general the overall objective is not to reduce the input while maintaining constant production but to attempt to maximize production with the resources available.

2.2 Measuring Effectiveness of Research Activities

Effectiveness of research activity can be estimated by means of two seemingly similar models that share a common feature: they account only for the output side and thus are acting as output aggregator functions. In the input side they rely on Koopman's idea of a henchman who has at his/her disposal a unitary quantity of an aggregated input. These two models are the BoD and the Kao and Hung (2003) (K&H) model, which gained increased popularity in recent years as models used to construct composite indicators. We next present and contrast these two models under three different specifications of (output) weights: variable, common and restricted.

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 unit, it does so by implicitly assigning less (more) weight to those sub-indicators or performance aspects that the assessed unit is a relatively weak (strong) compared to all other units in the sample. Moreover, the estimated weights are allowed to vary across units 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 actual to 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 attains by construction the maximum 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 unit. This in turn guarantees that any other weighting scheme would worsen the ranking of this unit. Moreover, when these weights are used by any other unit in the sample would not result to a composite indicator greater than one. The resulting weights are determined endogenously by solving for each evaluated unit the following problem:

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

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

where I_i^k is the i[th] sub-indicator of the k[th] unit, s_i^k are the weights to be estimated, *j* is used to index units and *i* to index sub-indicators which in our case correspond to different research outputs (i.e., types of publications, citations, patents).

The BoD model is equivalent to the multiplier form of the Charnes, Cooper and Rhodes (1978) input-oriented, constant returns to scale (CRS) DEA model when there is a single constant input that takes the value of one for all evaluated units.[8] Based on this, the dual formulation of the BoD model is given as:

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

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

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

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 constraints on the intensity variables it is clear that the BoD model exhibits constant returns to scale.[9]

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 research performance indicator but now under the assumption that this set of weights satisfies for each evaluated unit an adding-up/normalization constraint. The K&H model is written as:

$$E^k = \max_{u_i^k} \sum_{i=1}^{N} u_i^k I_i^k$$

---

[8] 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.

[9] 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 by working with the input-oriented formulation and in particular, its multiplier form.

$$st \sum_{i=1}^{N} u_i^k = 1$$

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

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 model. In addition, in the K&H model there is only one (equality) constraint, besides the non-negativity constraints of the weights, while in the BoD model the number of (inequality) constraints is equal to the number of evaluated units.

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 $u_i^k = \frac{s_i^k}{S}$. This implies that the K&H delivers values of the composite indicator that are close but not always equal to those imputed with the BoD model. More importantly, Karagiannis and Paschalidou (2014) note that while from the BoD weights we can derive the weights implied by the K&H model the opposite is not possible. This limitation of the K&H model is related to the type and number of constraints that involves. On the other hand, for this same reason, the K&H model is computationally less demanded. Lastly, at present the aggregation properties of the K&H model are not yet kwon and for this reason we cannot move the analysis of research productivity from the individual to institution (department or university) level in a theoretically consistent way.

In contrast, such an aggregation rule for the BoD model has been developed by Karagiannis (2013) within the framework of aggregate efficiency scores. In particular, it has been shown that the arithmetic average is the theoretically consistent aggregation rule for the BoD model; that is:

Thus, the aggregate composite performance indicator equals the simple (un-weighted) arithmetic average of the estimated individual composite indicators. This results from the single constant (unitary) input structure of the BoD model and the denominator rule (Fare and Karagianis, 2013) stating that consistency in aggregation of ratio-type performance measures, including efficiency indices, is ensured as long a the weights are defined in terms of the variable being in the denominator.[10] For an input-oriented model such as the BoD, these will be actual cost or input shares. But since all evaluated units have the same amount of (one unit) and face the same price for the single input, the share weights become equal to 1/K. In terms of research activity, this result implies that a department's research productivity can be simply estimated by means of the average research productivity of its faculty members.

Regarding now the estimation of research effectiveness in terms of common instead of variable weights, which according to Kao and Hung (2005) and Wang, Luo and Lin (2011) among others have the advantage of making it possible to compare and rank the performance of all evaluated units and not only classify them as efficient and inefficient, both the BoD and the K&H models poses special features.[11] First, for the BoD model Karagianis and Paleologou (2014) shown that common weights are related to average cross efficiency, which is of particular interest in assessing research productivity (Oral *et al.*, 2014) as it 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. In particular, average cross efficiency in the BoD model is based on a set of common weights given by the simple arithmetic average of weights obtained from the self-appraisal version of the model, i.e., the one discussed above. On the other hand, in the K&H model a set of common weights can be obtained by applying the compromise solution proposed by Kao and Hung (2005). In particular, we estimate a linear OLS (not including an intercept term) of the composite indicator obtained from the conventional form of the model on the set of sub-indicators under the restriction that the estimated parameters sum up to one.

The last set of weights on which a research productivity assessment may be carried out is that reflecting value judgment. For the BoD model, this is incorporated

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

[11] Another advantage of common weights is that they can be applied to calculate performance indices for DMUs not in the sample (Kao and Hung, 2007).

in terms of weights restrictions in the multiplier form of the model. Several types of weights restrictions have been used for this purpose including pie shares (see for example Cherchye *et al.*, 2007) and partial descending ordering (i.e., $s_1 1^{\uparrow} k \geq s_1 2^{\uparrow} k \geq s_1 3^{\uparrow} k \geq \cdots$). The latter is a particular interest case for the K&H model because then as Ng (2007, 2008) has shown there is no need to estimate the composite indicator by means of linear programming but rather to compute it based on partial averages; that is, the composite indicator is given as:

$$\max\left\{I_1^k, \left(\Sigma_{i=1}^{s} I_i^k\right)/_2, \left(\Sigma_{i=1}^{s} I_i^k\right)/_3, \ldots\right\}$$

where i is used to index sub-indicators.

Lastly, the BoD model, as a DEA-type model, can also be used to examine research productivity over time by means of the corresponding technology-based (i.e., Malmquist or Hicks-Moorsteen) indices, an aspect of performance evaluation that cannot be done with the K&H model. For the BoD model, Karagiannis and Lovell (2013) have shown that *first*, the Malmquist and Hicks-Moorsteen productivity indices coincide, they are multiplicatively complete, and the choice of orientation for the measurement of productivity change does not matter.[12] *Second*, there is a unique decomposition of the sources of productivity change containing three independent components, namely technical efficiency change, neutral technical change and output biased technical change. *Third*, the aggregate output-oriented Malmquist productivity index is given by the geometric average between any two periods of the simple (un-weighted) arithmetic average of the individual contemporaneous and mixed period distance functions.

3. An Application

The empirical application is from Karagiannis (2013) who apply the BoD to evaluate the research achievements of faculty members in the Department of Economics at the University of Macedonia, Greece during the period 2000-2006. In the proposed setting the single constant input corresponds to each faculty member and we consider

[12] A productivity index is multiplicatively complete if it can be written in a ratio form of input/output indices that are non-negative, non-decreasing, linearly homogenous scalar functions (O'Donnell, 2012).

two outputs, namely, journal articles and all other publications, which are measured by whole counting. As journal articles are considered all publications in outlets referenced in the Journal of Economic Literature and as other publications are considered papers published in journal not referenced in the Journal of Economic Literature, chapters in books and edited volumes. The relevant data reveal that on average each faculty member published almost one journal paper per year during the period 2000-2003 and it seems to be an improvement in research achievements as its annual average increase to somewhat above one during the period 2004 to 2006. The corresponding figures for other publications are well below one for the whole period, with a trend to decline significantly in the last two years. In addition, both kinds of publications are unevenly distributed between faculty members. There few faculty members with satisfactory achievements in journal article publications (more than two and a half on average per year) and one faculty member with similar performance in terms of other publications but most of them are around the departmental average. There were however two faculty members that had no journal article published during the whole period under consideration and with only one other publication each. Moreover, the achievements of newcomers are underestimated because of no entries in the data over the whole period under consideration. For convenience reason these two cases were disregarded and thus a total of 20 faculty members is included in the sample.

The average annual scores of technical efficiency, which reflect research efficiency at the department level, were found to be rather low and in the range of 0.36 to 0.43 indicating the relative heterogeneity in the achievements of faculty members. There are two peers with an average over time efficiency score of almost one and in addition, two other faculty members who are well above the average. However the great majority (16 faculty members) have an average efficiency score of less that 0.4, very close to departmental average. There are also three faculty member with comparatively low achievements and efficiency scores below 0.2. Nevertheless it seems that on the average there is a tendency for improvement in the degree of technical efficiency.

The productivity analysis shown that there was an annual increase of 6% in the research performance for the department as a whole. The main source behind this significant growth in publication rate was the improvement in technical efficiency. That is, the catching up process, namely the attempt of relatively inefficient faculty

members to catch up their peers and reach the frontier, proved to be the most important source of TFP growth. The second most important source was neutral technical change, which is associated with outward shifts of the frontier over time caused by the improvements in the performance of those faculty members that were operating almost efficiently during the years. The contribution of output biased technical change is very limited indicating most likely the presence of output neutral technical change. Nevertheless, there is a great variability in individual achievements regarding productivity: there were three faculty members with improvements well above the departmental average but there was on the other hand four faculty members that have deteriorated their performance during the period under consideration. This means that their publication record decreases over time. From the rest, there are two or three who kept their publication record unchanged over time. The great majority of faculty members however had shown small improvement in the range of 1-2% annually.

Figure 1: Average Annual Technical Efficiency

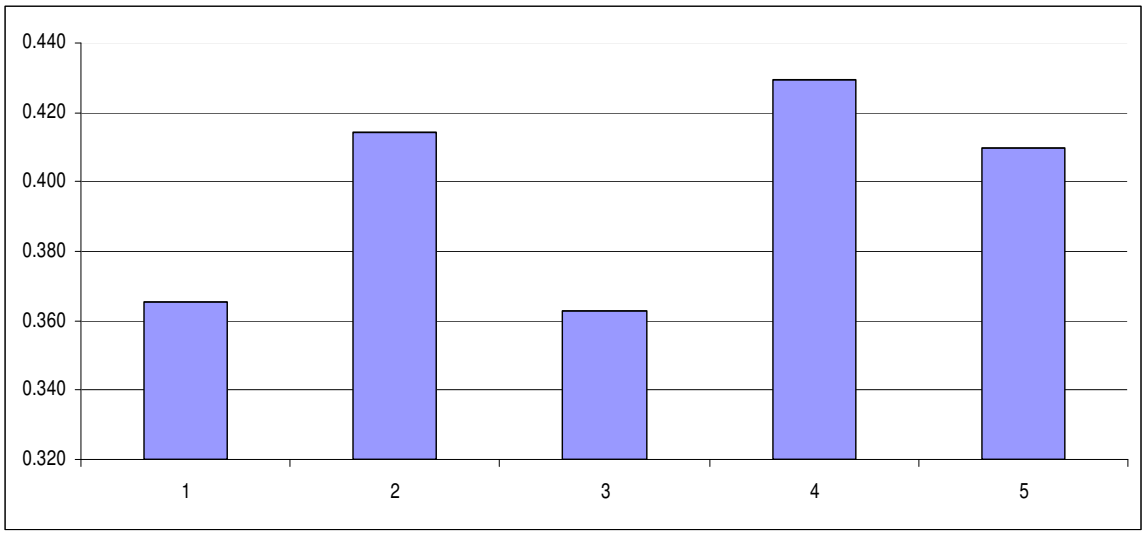
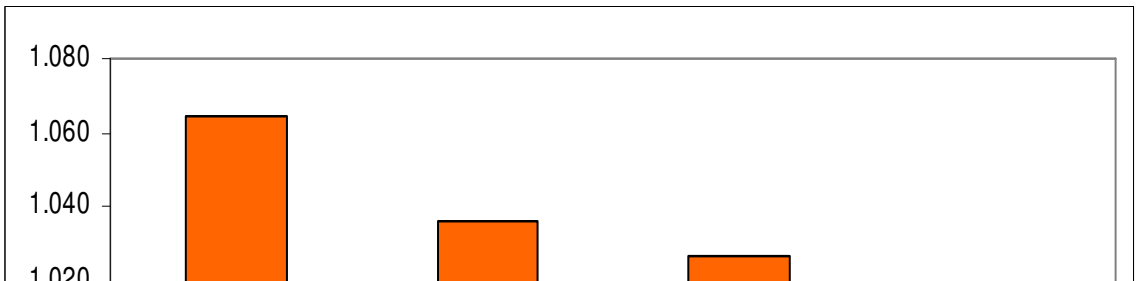


Figure 2: Sources of Productivity Growth, average annual values for the period 2000-2006



References

Abramo, G. and C.A. D'Angelo. How do you define and measure research productivity? *Scientometrics*, 2014 (forthcoming)

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, 1005-17.

Abramo, G., D'Angelo, C.A. and F. Di Costa. Citations versus journal impact factor as proxy of quality: Could the latter ever be preferable?, *Scientometrics*, 2010, 84, 821-33.

Abramo, G., D'Angelo, C.A. and F. Di Costa, Variability of research performance across disciplines within universities in non-competitive higher education systems, *Scientometrics*, 2014, 98, 777-95.

Abramo, G., D'Angelo, C.A. and F. Rosati. The importance of accounting for the number of co-authors and their order when assessing research performance at the individual level in life sciences, *Journal of Informetrics*, 2013, 7, 198-208.

Abramo, G., D'Angelo, C.A. and M. Solazzi. National research assessment exercises: A Measure of the distortion of performance rankings when labor input is treated as uniform, *Scientometrics*, 2010, 84, 605-19.

Abramo, G., Cicero, T. and C.A. D'Angelo. A sensitivity analysis of researchers' productivity rankings to the time of citation observation, *Journal of Informetrics*, 2012, 6, 192-201.

Abramo, G., Cicero, T. and C.A. D'Angelo. How important is the choice of the scaling factor in standardizing citations?, *Journal of Informetrics*, 2012, 6, 645-54.

Abramo, G., Cicero, T. and C.A. D'Angelo. Individual research performance: A proposal for comparing apples and oranges, *Journal of Informetrics*, 2013, 7, 528-39.

Caporaletti, L.E., Dula, J.H. and N.K. Womer. Performance Evaluation based on Multiple Attributes with Nonparametric Frontiers, *Omega*, 1999, 27, 637-45.

Charnes, A., Cooper, W.W. and E. Rhodes. Measuring the Efficiency of Decision Making Units, *European Journal of Operational Research*, 1978, 2, 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, 111-45.

Färe, R. and G. Karagiannis. The denominator rule for share-weighting aggregation. Unpublished manuscript, 2013.

Färe, R. and G. Karagiannis. A Postscript on Aggregate Farrell Efficiencies, *European Journal of Operational Research*, 2014, 233, 784-86.

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

Karagiannis, G. Faculty Research Performance and Efficiency Evaluation in Greek University Department of Economics Using DEA, paper presented in the *13th European Workshop on Efficiency and Productivity Analysis*, Helsinki, Finland, June 17-20, 2013.

Karagiannis, G. On Aggregating Cross Efficiencies, unpublished manuscript, 2014.

Karagiannis, G. On Structural and Average Technical Efficiencies, *Journal of Productivity Analysis*, 2015 (forthcoming).

Karagiannis, G. and C.A.K. Lovell. Productivity Measurement in Radial DEA Models with Multiple Constant Inputs, CEPA working paper, 2013.

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.

Karagiannis, G. and G. Paschalidou. Assessing Effectiveness of Research Activity at the faculty and the department level: A comparison of alternative models, paper presented in the *2nd Workshop on Education Efficiency*, London, Sept 26-29, 2014.

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

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

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

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, 613-28.

Lin, C.S., Huang, M.H. and D.Z. Chen. The influences of counting methods on university rankings based on paper count and citation count, *Journal of Informetrics*, 2013, 7, 611-21.

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, 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, 46-51.

Ng, W.L. A Simple Classifier for Multiple Criteria ABC Analysis, *European Journal of Operational Research*, 2007, 177, 344-353.

Ng, W.L. An Efficient and Simple Model for Multiple Criteria Supplier Selection Problem, *European Journal of Operational Research*, 2008, 186, 1059-67.

O'Donnell, C. J. An Aggregate Quantity Framework for Measuring and Decomposing Productivity Change, *Journal of Productivity Analysis*, 2012, 38, 255-72.

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, 20-28.

Wang, Y.M., Luo, Y., & Lan, Y.X. (2011). Common weights for fully ranking decision making units by regression analysis, *Expert Systems with Applications*, 38, 9122-9128.