Scientometrics (2010) 85:657–680
DOI 10.1007/s11192-010-0286-5

# To publish or not to publish? On the aggregation and drivers of research performance

**Kristof De Witte · Nicky Rogge**

Received: 15 December 2009 / Published online: 5 September 2010


**Abstract** This paper presents a methodology to aggregate multidimensional research output. Using a tailored version of the non-parametric Data Envelopment Analysis model, we account for the large heterogeneity in research output and the individual researcher preferences by endogenously weighting the various output dimensions. The approach offers three important advantages compared to the traditional approaches: (1) flexibility in the aggregation of different research outputs into an overall evaluation score; (2) a reduction of the impact of measurement errors and a-typical observations; and (3) a correction for the influences of a wide variety of factors outside the evaluated researcher's control. As a result, research evaluations are more effective representations of actual research performance. The methodology is illustrated on a data set of all faculty members at a large polytechnic university in Belgium. The sample includes questionnaire items on the motivation and perception of the researcher. This allows us to explore whether motivation and background characteristics (such as age, gender, retention, etc.,) of the researchers explain variations in measured research performance.

**Keywords** Research performance · Data envelopment analysis · Conditional efficiency · Higher education · Composite indicator

**JEL classification** C14 · C25 · I21

K. De Witte
TIER, Faculty of Economics and Business, Maastricht University,
Kapoenstraat 2, 6200, MD, Maastricht, the Netherlands
e-mail: Kristof.dewitte@econ.kuleuven.be

K. De Witte · N. Rogge
Faculty of Business and Economics, Katholieke Universiteit Leuven (KULeuven),
Naamsestraat 69, 3000 Leuven, Belgium

N. Rogge (✉)
Centre for Economics and Management (CEM), Hogeschool-Universiteit Brussel (HUBrussel),
Stormstraat 2, 1000 Brussels, Belgium
e-mail: Nicky.Rogge@hubrussel.be

## Introduction

Universities and colleges are increasingly interested in evaluating the performances of their academic staff, both in terms of teaching performance and of research performance. Given the growing attention to research, this paper will focus exclusively on research performance. In particular, this paper presents (1) a flexible tool to evaluate the multiple dimensions of research performance and (2) relates research performance to individual characteristics, motivation, and employment conditions.

Current literature on research evaluation mainly employs single-criterion measures, such as reputational ratings gathered by polls or peer reviews, number of publications (eventually in conjunction with a journal quality index or a citation index, e.g., Van Leeuwen et al. 2003; van Raan 1996) in a predefined set of refereed journals (e.g., Zamarripa 1995; Sax et al. 2002), or citation counts (e.g., McCain and Turner 1989; Nederhof et al. 1993; Lee and Bozeman 2005). Recently, several opponents have criticized such simplistic measures doubting whether they are able to accurately convey research performance. In their opinion, the nature of research is by far too complex to be grasped by one single output criterion. For instance, Martin (1996) note that few would dispute that research is multi-dimensional in terms of its nature and outputs. Avital and Collopy (2001) agree arguing that research performance is broader than only one-dimensional measures. They suggest that (compared to single-criterion measures) multidimensional instruments are less sensitive to systematic measurement error and biases created by researchers who adjust their behavior in an attempt to improve evaluations (Avital and Collopy 2001, p. 53). Hattie and Marsh (1996) also argued that weighted measures of research performance may be preferable to single-criteria numbers.[1] In view of these considerations, a multi-criteria measure seems more appropriate in assessments of researchers' performances.

However, the construction of a multi-criteria Research Evaluation Score (RES-score) is an intricate matter with, amongst others, two important conceptual and methodological difficulties to overcome[2]: (1) How should one weight and aggregate the different output criteria? Or, stated differently, how important are the several research outputs in the overall performance evaluation? Is it legitimate to assign a uniform set of weights over the several output criteria (i.e., equal/fixed weights)? Also, is it legitimate to apply a uniform set of weights to all evaluated researchers? Some researchers are clearly specializing in writing international books, while other are specializing in attracting research funding. Using the same weights for all researchers, would be considered as unfair within a research unit.[3] (2) How should the RES-scores be adjusted for the impact of exogenous characteristics which are (often) beyond the control of the researcher? There are numerous findings in the academic literature which suggest that some background characteristics (e.g., age, gender,

[1] Note that the use of multidimensional measures has not only been suggested for evaluating individual researcher performances, but also for evaluating research departments. Vinkler (1998, 2006), Ruiz et al. (2010), and Bonaccorsi et al. (2006), for instance, developed composite (i.e., multidimensional) indicators for evaluating the research performances of research institutes.

[2] Another conceptual difficulty is the choice of academic output criteria that are deemed appropriate to be present in the performance evaluation. Selecting the relevant output criteria is the duty of faculty board members, the evaluated researchers and experts in evaluation methods (primarily scientometricians).

[3] The question is even more prominent in the application at hand (see below). Similar to 'new' (polytechnic) universities in the UK and the colleges in the US, objectives of researchers in the application are more diverse than in 'traditional' research departments (although some diversity might be present there as well). Some researchers are specializing in writing international books, others in attracting research funding.

rank/tenure, time spent on teaching, department policy, etc.,) may have a significant impact on the research performance of academic staff (Bellas and Toutkoushan 1999; Hattie and Marsh 1996, 2002; Lee and Bozeman 2005; Maske et al. 2003; Ramsden 1994; Sax et al. 2002; Chen et al. 2006). Intuitively, researchers realize some conditions are more beneficial to productive research while other conditions are more detrimental. Yet, traditional RES-scores do not account for differences in these uncontrollable conditions. Consequently, these scores are inherently biased towards researchers working under more favorable conditions. With this 'bias' in mind, several practitioners and researchers have claimed that uncorrected scores tend to be unfair as they give an advantage to those who work in more constructive conditions. The opposite reasoning holds true for academics who work under less favorable conditions. In their case, it is more difficult to obtain a good performance level (and, hence, a good RES-score). Thus, the remark of Emery et al. (2003, p. 44) made with respect to teacher evaluation tools, also applies to evaluation instruments for faculty research performances: "Any system of faculty evaluation needs to be concerned about fairness, which often translates into a concern about comparability. Using the same evaluation system [without properly accounting for the differences in research conditions] for everyone almost guarantees that it will be unfair to everyone". Stated differently, unadjusted RES-scores are potentially flawed and, therefore, unreliable as a measure of researcher performance. However, we are unaware of any study which corrects RES-scores for heterogeneity in (potentially) influential characteristics and conditions not under the control of the evaluated researchers.[4]

The contributions of this paper are threefold. A first contribution of this paper is the proposal of a global RES-score. This study adds to the extant literature by outlining the weighting issue in the construction of a composite RES-score. In Section "The weighting issue" and "Methodology", we advocate a methodology to construct RES-scores which does address the weighting and correcting issues. In particular, we suggest a specially tailored version of the non-parametric Data Envelopment Analysis model (DEA; Charnes et al. 1978). The so-called Benefit of the Doubt model (BoD) allows for the aggregation of various dimensions of research performance while incorporating the relative importance of these dimensions (e.g., a publication in an A-journal is more valued than a B-journal). The core idea is that output criteria on which the evaluated researcher performs well compared to his/her colleagues, should weight more heavily than the output criteria on which he performs relatively poor. The rationale for doing so is that a good (poor) relative performance is considered to be an indication of a high (low) attached importance by the evaluated researchers. Similar to 'new' (polytechnic) universities in the UK and the colleges in the US, researchers in our application are more diverse than in 'traditional' research departments (although large diversity might be present there as well). Some researchers are, e.g., specializing in writing international books, others in attracting research funding. The BoD model accounts for this by endogenously weighting the research outputs. Using endogenous weights, the 'benefit-of-the-doubt' (BoD) model allows each researcher for a certain degree of specialization. As such, it avoids the subjectivity of fixed weights.

[4] Note that if research is considered as an 'absolute competition' (see, e.g., Merton 1968; Mercer and Wanderer 1970), one can argue that there is no need to account for background characteristics as age or teaching load. The proposed model can be easily adapted to neglect exogenous conditions. If research is considered as a 'relative competition' among faculty members (e.g., for in personnel decisions), exogenous background should be accounted for.

The BoD has already been extensively described and studied in the JRC-OECD Handbook on Constructing Composite Indicators (OECD 2008). It has also been used to construct composite indicators in wide ranging fields such as the economy (e.g., the Internal Market Index; Cherchye et al. 2007a), human development (e.g., the Human Development Index; Despotis 2005), technological development (e.g., the Technology Achievement Index; Cherchye et al. 2008), creative economy (Bowen et al. 2008), and sustainable development (e.g., Sustainable Development Index; Cherchye and Kuosmanen 2006). Recently, the European Commission has used the BoD technique to gauge member states' performance with regard to the Lisbon objectives (European Commission 2004, pp. 376–378).[5]

Secondly, this paper attempts to fill the gap in (1) estimating the impact (in both size and direction) of background conditions on the measured research performances, and (2) correcting the RES-scores for the (un)favorable conditions in which the researcher works. In particular, we examine whether productivity in research can be related to a set of items describing individual researcher motivations and perceptions with respect to teaching and research (as well as the nexus between both activities), personal characteristics (e.g., age, gender), and working conditions (e.g., retention, teaching load, and time for research). From the point of view of university management, both types of information are useful. For instance, evaluation scores and rankings are particularly helpful in personnel decisions (e.g., recruitment, reappointment, promotion, retention, and dismissal, etc.,). The explanatory information, on the other hand, provides insights on the exact impact of working conditions on research performance can guide university management in attempts to facilitate an environment that is more conducive to creativity and productive research.[6]

Thirdly, to illustrate the practical usefulness of the approach, we apply the model on a dataset collected at department 'Business Administration' of the Hogeschool Universiteit Brussel (Belgium) in the academic years 2006–2007 and 2007–2008. This university college resembles in many ways to the 'new' (polytechnic) universities in the UK and the colleges in the US. In particular, it used to be an educational institution with exclusive focus on teaching, but recently, thanks to the Bologna reforms (and an academization process initiated by the Flemish Government), it became increasingly research-oriented. The large resemblance with higher education institutions in other countries implicates that the university college under study is an excellent example to illustrate the usefulness of the presented "Methodology". The data set comprises output (research) data on all 81 researchers. We matched this data set with administrative and survey data. The administrative data contains information on age, gender, doctoral degree, tenure, (official) teaching load, and (official) time for research. The data are further enriched with a questionnaire on the researcher's opinions and perceptions on research satisfaction and personal goals.

The remainder of the paper is organized as follows. While the next section discusses the "The weighting issue" and the advantages of our methodology, in "Methodology" section we present the basic DEA model as well as its robust (extreme observations and/or data

[5] Other approaches to construct composite indicators of complex phenomena are also discussed in the JRC-OECD Handbook (and in other reports such as Business Climate Indicators (DG ECFIN), Economic Sentiment Indicators (EU), Composite Leading Indicators (OECD), General Indicator for Science and Technology (NISTEP, Japan), or Composite Index of Technological Capabilities (Archibugi and Coco 2004).

[6] Nevertheless, the methodology does not examine the potential reverse causality among the variables (e.g., time for research may be endogenous to research output). Examining the causality of the variables requires besides the use of instrumental variables, a less flexible parametric framework. We consider this as scope for further research.

measurement errors) and conditional (heterogeneity among researchers) extensions; while the subsequent sections report "The data" and the "Results". In the final section, we offer some concluding remarks and some avenues for further research.

## The weighting issue

The few studies which use multi-criteria instruments, calculate commonly the global RES-score as an arithmetic mean or a weighted sum of the researchers' performances on the several output criteria:

$$\mathrm{RES}_c = \sum_{i=1}^{q} w_i y_{c,i}, \tag{1}$$

where $y_{c,i}$ is the number of publications the evaluated researcher $c$ realized in the research output category $i$; $w_i$ the importance weight assigned to the publications pertaining to the output category $i$ (with $0 \leq w_i \leq 1$ and $\sum_{i=1}^{q} w_i = 1$); $q$ the number of output criteria considered in the research evaluation. In studies where the RES-scores are computed as an arithmetic mean: $w_i = 1/q$.[7] This implies that all aspects of research are assumed to be of equal importance. In essence, an arithmetic mean RES-score corresponds to a single-criteria measure where the publications are just counted over the different research output categories without any correction for the quality. When the RES-score is constructed as a weighted sum of publications with $w_i$ varying over the different research output categories, this score corresponds essentially to a simple publication count with a correction for quality (e.g., Kyvik 1990).

In both cases, weights are uniform for all evaluated researchers. Moreover, when using an arithmetic average, weights are even uniform over the several output criteria. Whether such uniform weights (over output criteria and/or for evaluated researchers) are legitimate is questionable. There are some indications suggesting that uniformity of weights across research criteria and/or over researchers is undesirably restrictive. Among others, Massy and Wilger (1995) assert that some accounting for quality differences over output criteria is necessary in any definition of research performance.

However, defining accurate importance values $w_i$ for the different output criteria is a very difficult task. First of all, there is a lot of diversity among the beliefs held by many academic administrators and faculty researchers about what are correct weights for the different output criteria. This makes it difficult to come to an agreement on the relative weights to be attached to each type of publication. Any choice of fixed weights will be subjective to some extent. Further, given that varying weights may result in varied RES-scores, and shift the rankings as a result, this subjectivity in weight choice is very likely to be interpreted as unfair. Unsurprisingly, disappointed researchers will invoke this unfairness and the subjectivity in weight choice to undermine the credibility of the RES-scores. A potential solution to this concern is to allow a limited amount of variation in the aggregation weights over the researchers. The question then arises: what is the amount of variation that is allowed?

[7] In this paper, the labels 'weights', 'importance weights', and 'importance values' are used interchangeable, thereby referring to the value that is attached to the research output criteria in the development of the global RES-score.

To account for the weighting issues in the construction of the research evaluation scores, this paper proposes a specially tailored version of the Data Envelopment Analysis methodology (DEA; the model is outlined in the "Methodology").[8] The basic DEA model has been developed by Charnes et al. (1978) as a non-parametric (i.e., it does not assume any a priori assumption on the production frontier) technique to estimate efficiency of observations. Here, we do not apply the original DEA model, but rather an adapted approach which originates from DEA. This so-called BoD model exploits a key feature of DEA. In particular, thanks to its linear programming approach, DEA allows for an endogenous weighting of multiple output/performance criteria (Melyn and Moesen 1991).[9] This data-driven weighting procedure has five important advantages compared to the traditional model as in Eq. 1.

Firstly, for each evaluated researcher, weights for the various output criteria are chosen such that the most favorable RES-score is realized. One could intuitively argue that, given the uncertainty and lack of consensus on the true weights of research outputs, BoD looks for those weights which put the evaluated researcher in the best possible light compared to his/her colleagues. As such, research performance is considered as a relative standard. Similar to all performance estimations, research performance is a relative issue which depends on the reference sample. The BoD model grants the 'benefit-of-the-doubt' to each researcher in an already sensitive evaluation environment. Being evaluated optimally, disappointed researchers (i.e., researchers with RES-scores below expectations) can not longer blame these poor evaluations to subjective or unfair weights. Any other weighting scheme than the one specified by the BoD model would worsen the RES-score. Secondly, the BoD model is flexible to incorporate stakeholder opinion (e.g., researchers, faculty administrators, and experts) in the construction of the RES-scores through pre-specified weight restrictions, to ensure that importance values are chosen in line with 'agreed judgments' of these stakeholders.[10] Particularly with an eye towards evaluations of research personnel, this advantage is essential for the credibility and acceptance of RES-scores. Massy and Wilger (1995) emphasized that some considerations for quality differences over output criteria are indispensable in a reasonable research evaluation. Thirdly, researchers are evaluated relative to the observed performances of colleagues. This clearly marks a deviation from the common practice in which benchmarks are exogenously determined by department administrators often without any sound foundation. Fourthly, we can adjust the BoD model such that its outcomes are less sensitive to influences of outlying or extreme observations as well as potential measurement error in the data. In particular, we apply insights from the robust order-$m$ efficiency scores of Cazals et al. (2002) to our specific BoD setting. Finally, the BoD model can be adjusted (after the conditional efficiency approach of Daraio and Simar 2005, 2007a, b) to account for background influences (e.g., age, gender, rank, PhD, teaching load, time for research, etc.,). In our case of research

[8] Although the DEA model has not been applied in the construction of RES-scores, the literature counts various studies focusing on the efficiency in research activities of universities or (university or research) departments (e.g., Johnes and Johnes 1993; Beasly 1995; Bonaccorsi et al. 2006; Ruiz et al. 2010; Cherchye and Vanden Abeele 2005).

[9] In previous studies, Rogge (2009a, b) and De Witte and Rogge (2009) proposed a similar 'Benefit-of-the-Doubt' variant of DEA to construct teacher evaluation scores (SET-scores) based on student questionnaire data on multiple underlying performance indicators (measuring several aspects of teacher performance). Particularly the construction of weight restrictions differ across the issues at stake.

[10] This reasoning is very much in line with the remark of Foster and Sen (1997, p. 206) that while it is difficult to let stakeholders agree on a unique set of weights, it is easier to let them agree on restrictions on these weights.

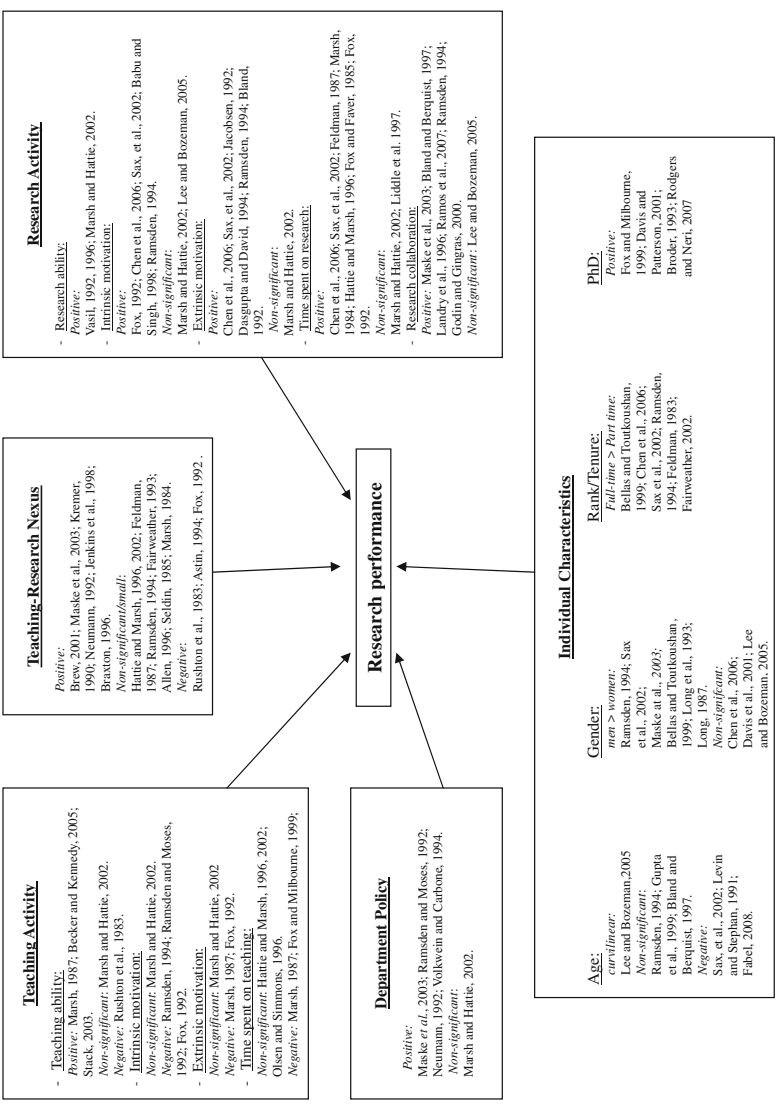


**Fig. 1** Researcher characteristics, motivations, and working conditions

evaluations, this technique enables us to include different (potentially) influential conditions (outside the control of the researcher) into the built-up of the global RES-scores. Several studies (e.g., Hattie and Marsh 1996, 2002; Harris and Kaine 1994) have focused on the impact of such characteristics, opinions and perceptions on research performances. As Fig. 1 shows, results are rather mixed. The size and direction of the associations seem to be dependent on the circumstances, the content, the specificities of the considered evaluation instrument (i.e., single-criteria vs. multi-criteria measure), and the methodology used to examine the relationships (e.g., multilevel modeling vs. regression analysis). Using a nonparametric technique, we try to limit the a priori assumptions and, as such, try to obtain more reliable estimates of the importance of the variables.

## Methodology

### The 'benefit-of-the-doubt' model

"The weighting issue" section discussed some advantages of the proposed BoD model. The BoD model relies on the non-parametric DEA approach, which is an efficiency measurement technique originally developed by Farrell (1957) and put into practice by Charnes et al. (1978). In essence, DEA is a non-parametric model and, hence, does not require any a priori knowledge on the 'functional form' of the production function (i.e., the production function underlying the studied phenomenon). Obviously, this non-parametric feature is important in the evaluation of complex phenomena where objective knowledge on the underlying structure is usually lacking. In comparison to the traditional DEA-problem, the only difference is that the development of an overall RES-score only requires a look at the individual performances of researchers in the different research criteria $i$ (with $i = 1, \ldots, q$) (thus, considering the outputs without explicitly taking into account the input dimension).[11] In this latter context, Melyn and Moesen (1991) alternatively labelled this method as the 'benefit-of-the-Doubt'-approach, a label that originates from one of the remarkable features of DEA: information on the appropriate weights can be retrieved from the observed data themselves (i.e., endogenous weighting).

Specifically, the core idea is that output criteria on which the evaluated researcher performs well compared to his/her colleagues in the reference set $\Upsilon$ should weight more heavily than the output criteria on which he/she performs relatively poor.[12] The rationale for doing so is that a good (poor) relative performance is considered to be an indication of a high (low) attached importance by the evaluated researchers.[13] For example, if, in comparison to his/her colleagues (i.e., all observations $y_{j,i}$ in the reference set $\Upsilon$), the researcher under evaluation published a high number of papers in international journals this reveals

[11] As Cherchye et al. (2007b, p. 121) pointed out, this BoD model is formally tantamount to the original input-oriented CCR-DEA model (Charnes et al. 1978), with all research output criteria $q$ considered as outputs and a 'dummy input' equal to unity for all the researchers. An intuitive interpretation may be obtained simply by regarding the BoD-model as a tool for aggregating performance on multiple output criteria (without explicit reference to the inputs). For a more elaborate discussion of the BoD-model and its intuitive interpretation, we refer to Cherchye et al. (2007b), Lovell et al. (1995) and Cook (2004).

[12] Recently, Claro and Costa (2010) presented another possible performance indicator in which a researcher's output is compared to a reference set of research output from top researchers (hence, a relative perspective in the evaluation).

[13] The advanced specialization of new 'polytechnic' universities and colleges obliges universities to allow for specialization in research output criteria. The BoD model accounts for diverging specializations.

that the researcher considers such publication to be of high importance. Consequently, his/her performances should weight more heavily on this criterion (i.e., high weight $w_{c,i}$). In other words, for each researcher separately, BoD looks for the weights that maximize (minimize) the impact of the criteria where the researcher performs relative good (poor) compared to the other researchers. Hence, BoD-weights $w_{c,i}$ are optimal and yield the maximal RES-score.[14] This gives the following linear programming problem for each researcher under consideration $c$:

$$\mathrm{RES}_c(y) = \max_{w_{c,i}} \sum_{i=1}^{q} w_{c,i} y_{c,i} \tag{2}$$

s.t.

$$\sum_{i=1}^{q} w_{c,i} y_{j,i} \leq 1 \quad j = 1, \ldots, c, \ldots, n \tag{2a}$$

$$w_{c,i} \geq 0 \quad i = 1, \ldots, q. \tag{2b}$$

The objective function (2) reveals the 'benefit-of-the-doubt' interpretation: the BoD model lets the data speak for themselves and endogenously selects those weights $w_{c,i}$ which maximize the RES-scores. Any other weighting scheme than the one specified by the BoD model would worsen $\mathrm{RES}_c(y)$. This data-orientation is justifiable in the context of evaluating research performances where there is usually a lack of agreement among stakeholders (i.e., policy makers, researchers, etc.,), and uncertainty about the proper importance values of the research output criteria. This perspective clearly deviates from the current practices of using single-criterion measures or multiple-criteria as in (1) with or without a correction for the perceived quality.

Notice that the standard BoD model as in (2)–(2b) grants evaluated researchers considerable leeway in the definition of their most favourable weights $w_{c,i}$. More precisely, only two (rather minimal) constraints have to be satisfied. A first one is the 'normalization' constraint (2a) that ensures that all RES-scores computed with the evaluated researcher's most favourable weights $w_{c,i}$, can at most be unity (or, equivalently, 100%). Thus, we obtain $0 \leq \mathrm{RES}_j \leq 1$ $(j = 1, \ldots, c, \ldots, n)$ with higher values indicating better overall relative research performances. The second 'non-negativity' constraint limits weights to be non-negative (hence, $w_{c,i} \geq 0$). Apart from these restrictions (2a) and (2b), weights can be chosen completely free to maximize the RES-score of the evaluated researcher vis-à-vis the other researchers. However, in some situations, it can allow a researcher to appear as a brilliant performer in a way that is difficult to justify. For instance, while having publications in several research output criteria, some researchers may prefer to only consider one of these criteria (i.e., the one in which the researcher performs best relative to their colleagues) in the built-up of their RES-scores (thus, assigning zero-weights to all other criteria) without violating the two basic restrictions. In such research evaluations, global RES-score reduce to the researchers' performances on one single dimension. Another concern is that chosen BoD-weights may too much deviate from what stakeholders (i.e., the faculty board, evaluated academics) believe is appropriate. Without doubt, opponents of research evaluations will claim that RES-scores based on improper weights are not meaningful.

[14] For completeness, we mention that BoD alternatively allows for a 'worst-case' perspective in which entities receive their worst set of weights, hence, high (low) weights on performance indicators on which they perform relative weak (strong) (Zhou et al. 2007).

**Table 1** Criteria for the evaluation of research performance

| No. | Research outlet | Maximum weight |
|---|---|---|
| 1 | International publication (Thompson Master List) | 15 |
| 2 | International book (based on own scientific work) as an author | 15 |
| 3 | National book (based on own scientific work) as an author | 10 |
| 4 | Finished research report (externally funded by a commissioner) | 10 |
| 5 | National scientific journals, chapter in international scientific book, Complete article in international proceedings (in all cases peer-reviewed) | 8 |
| 6 | Promoter of an externally funded project | 5 |
| 7 | Complete article in national proceedings (peer-reviewed) | 7 |
| 8 | Chapter in a scientific national book | 7 |
| 9 | Research/discussion paper in HUB or other series | 5 |

Fortunately, BoD models are flexible to incorporate additional restrictions. Formally, this involves adding the general weight constraint (2c) to the standard BoD model:

$$w_{c,i} \in W_e \quad i = 1, \ldots, q \text{ and } e \in E \tag{2c}$$

with $W_e$ denoting the set of permissible weight values defined based upon the opinion of selected stakeholders $e \in E$. Especially with an eye towards practical evaluations, it is crucial for the credibility and acceptance of RES-scores to define such weight restrictions based on stakeholder opinions (if available). Formally, the complete ordinal ranking of the importance values of the nine research output criteria, as agreed upon by the stakeholders (see below and Table 1), is presented as follows:

$$w_{c,1} = w_{c,2} \geq w_{c,3} = w_{c,4} \geq w_{c,5} \geq w_{c,7} = w_{c,8} \geq w_{c,6} = w_{c,9} \geq 0.01 \tag{3}$$

From a technical perspective, we have to adjust these additional weight restrictions for the potential presence of zero values in the evaluation data. Indeed, in one or multiple output dimensions researchers may not have been able to produce any publication during the evaluation period (hence, the associated $y_{c,i}$'s are equal to zero). The endogenous weighting procedure of BoD will automatically assign a zero weight to such output criteria. However, in our evaluation procedure (with the additional ordinal weight restrictions as specified above), this standard procedure may lead to infeasibilities. Kuosmanen (2002) and Cherchye and Kuosmanen (2006) proposed a simple modification of the weight restriction to prevent this infeasibility: multiply the constraints by the product of the corresponding $y_{c,i}$'s.[15] Formally,

$$\begin{aligned} &\left(w_{c,1} - w_{c,2}\right) \times y_{c,1} \times y_{c,2} = 0 \\ &\left(w_{c,2} - w_{c,3}\right) \times y_{c,2} \times y_{c,3} \geq 0 \\ &\ldots \\ &\left(w_{c,6} - w_{c,9}\right) \times y_{c,6} \times y_{c,9} = 0 \\ &w_{c,i} \times y_{c,i} \geq y_{c,i} \times 0.01 \quad \forall i = 1, \ldots, 9. \end{aligned} \tag{4}$$

In this adjusted version of the additional weight restrictions, a standard weight $w_{c,i} = 0$ for an output criterion $i$ with $y_{c,i} = 0$ does no longer enforce other weights to be equal to zero. In cases where one or both of the associated $y_{c,i}$'s equals zero, the restriction becomes

[15] See Kuosmanen (2002) for a more comprehensive discussion.

redundant and hence has no further influence on the other restrictions in (4). If none of the associated $y_{c,i}$'s are equal zero, then the adjusted version of the weight restriction reduces to the original restriction as in (3). Formally, the introduction of these adjusted weight restrictions the basic BoD model entails replacing the general weight constraint (2c) for (4).

## The robust BoD model

Similar to other nonparametric techniques, in its basic form, BoD suffers from a sensitivity to the influences of outliers, extreme values, and other data irregularities (e.g., measurement errors). This sensitivity results from: (1) the deterministic nature of BoD by which all differences between the performances of the evaluated researcher $y_{c,i}$ and the other research performances $y_{j,i}$ in the reference set $\Upsilon$, are perceived as a perfect reflection of actual differences in research performance, and (2) the modeling assumption that all n observations (thus also potential outliers or observations infected by measurement errors) should be included in the reference set $\Upsilon$ (see constraint (2a)). Because of these two assumptions, the presence of only one atypical/extreme research performance in the reference set $\Upsilon$ suffices to alter the RES-scores of all researchers dramatically.[16]

The robust order-$m$ methodology of Cazals et al. (2002) allows overcoming this aforementioned limitation.[17] Basically, this robust approach no longer puts central the traditional assumption that all observations should be considered in the computation of the RES-scores. Instead, using a simple Monte Carlo simulation technique, one draws repeatedly (i.e., $B$ times) and with replacement $m$ observations from the original reference set $\Upsilon$ of $n$ observations.[18] This smaller reference set is labelled $\Upsilon_c^{b,m}$ (with $b = 1, \ldots, B$). For each of the $B$ draws, the BoD-based RES-scores are computed relative to this sub sample of size $m$. By taking subsamples, the robust order-$m$ technique reduces the impact of outlying observations.

$$\mathrm{RES}_c^{b,m}(y) = \max_{w_{c,i}} \sum_{i=1}^{q} w_{c,i}^b y_{c,i} \tag{5}$$

$$s.t.$$

$$\sum_{i=1}^{q} w_{c,i}^b y_{j,i} \leq 1 \quad j = 1, \ldots, m \quad \forall y_{j,i} (j = 1, \ldots, m) \in \Upsilon_c^{b,m} \tag{5a}$$

$$w_{c,i}^b \geq 0 \quad i = 1, \ldots, q \tag{5b}$$

$$w_{c,i}^b \in W_e \quad i = 1, \ldots, q \text{ and } e \in E. \tag{5c}$$

Formally, the robust BoD model as in (5)–(5c) is largely similar to the original BoD model as in (2)–(2c). In fact, the only difference is situated in the composition of the

[16] In our data set, for instance, there is one researcher 'k' who succeeded to publish 27 research reports (externally funded by a commissioner) in the period under study (i.e., $y_{k,4} = 27$). At first sight, this seems to be an example of an outstanding research performance. However, a more profound analysis of this figure indicated that all research reports were part of one major project in which a particular study was made for 27 municipalities. The result of this study was summarized in reports for each municipality separately.

[17] An alternative to the order-m approach of Cazals et al. (2002) is the order-$\alpha$ approach of Daouia and Simar (2007). The ideas behind both techniques are largely similar. In fact, the adjustment of the order-m ideas to the order-$\alpha$ ideas, and vice versa, is straightforward (see, Daraio and Simar 2007a, pp. 65–76).

[18] Note that a particular research performance can be drawn multiple times in the same Monte Carlo step.

reference set. Further note $w_{c,i}^{b}$ instead of $w_{c,i}$ as optimal weights are now computed $B$ times. Recall that the general weight constraint (5c) represents the adjusted ordinal weight restriction as in (4). Having obtained the B RES-scores, we compute the robust version of $\text{RES}_c(y)$, $\text{RES}_c^m(y)$, as the arithmetic average of the $B$ $\text{SET}_c^{b,m}(y)$ estimates:

$$\text{RES}_c^m(y) = \frac{1}{B}\sum_{b=1}^{B}\text{RES}_c^{b,m}(y). \tag{6}$$

Besides mitigating the impact of outlying observations, Jeong et al. (2010) show that the order-$m$ estimates have additional attractive properties in that they are consistent and have a fast rate of convergence.[19]

In contrast to the traditional BoD estimates, the robust $\text{RES}_c^m(y)$ scores can be larger than unity. Indeed, thanks to drawing a subsample of $m$ observations with replacement from the full sample $\Upsilon$ the evaluated research performance $c$ can be compared with a reference sample $\Upsilon_c^{b,m}$ consisting of researchers with, on average, a lower performance level. As such, outstanding research performances (i.e., observations with a $\text{RES}_c^m(y) > 1$) could arise. A resulting $\text{RES}_c^m(y) = 1$ indicates that the evaluated researcher $c$ performs on a level that is similar to the average performance level realized by expected $m$ peers. Finally, a $\text{RES}_c^m(y) < 1$ points to a research performance that is worse compared to the average order-$m$ benchmark research performance.[20]

The robust and conditional BoD model

As discussed before, background characteristics $z$ may play a role in research performance (see Fig. 1). We account for these (often, but not always) exogenous characteristics by applying insights from the conditional DEA model to the BoD model. The former model has been proposed by Cazals et al. (2002) and Daraio and Simar (2005) and further extended by Daraio and Simar (2007a, b) to multivariate (continuous) characteristics, by Badin et al. (2010) to an improved bandwidth estimator and by De Witte and Kortelainen (2008) to discrete characteristics.

The conditional efficiency approach extends the robust order-$m$ model of Cazals et al. (2002) by drawing the $m$ observations with a particular probability (instead of drawing at random). The probability is obtained from estimating a non-parametric kernel around the background characteristics $z$ of the evaluated observation (we estimate a kernel as this allows us to smooth the background variables). As such, only observations which have similar background characteristics enter the reference group against which relative performance is estimated. Algebraically, model (5) is altered by restricting the reference set $\Upsilon_c^{b,m}$ to $\Upsilon_c^{b,m,\tilde{z}}$, where $\tilde{z}$ denotes observations which have similar background characteristics as $z$. The obtained $\text{RES}_c^m(y|z)$-score properly accounts for the background characteristics.

The mixed kernel smoothing of De Witte and Kortelainen (2008), which applies the mixed kernels from Li and Racine (2007), conveniently accounts for insignificant background characteristics by oversmoothing the kernel. In particular, if multiple background characteristics are included in the analysis, from which some turn out to have an

[19] Although these attractive properties were derived for the original DEA model, the extension to the BoD approach is rather straightforward.

[20] We follow Daraio and Simar (2005, 2007a, b) in selecting the size of the subsample $m$ as the value for which the percentage of super-efficient observations (i.e., RES > 1) becomes relatively stable. In our particular application, $m$ is determined as $m = 40$ (although sensitivity analysis with different values of $m$ shows the robustness of the approach).

insignificant impact on the RES-scores, the kernel bandwidth becomes very large in the insignificant dimension such that the insignificant variable becomes irrelevant for the computation of the conditional $RES_c^m(y|z)$-score (see Li and Racine 2007). This is convenient as, therefore, no a priori assumptions on the influence and direction of the background characteristics have to be made.

Besides the outlined advantages of the robust BoD model, the conditional model has two additional advantages. Firstly, as discussed in Daraio and Simar (2005), the fully nonparametric model does not impose a separability condition between the output variables and the background characteristics. In other words, the model acknowledges that background characteristics (as the ones presented in Fig. 1) may influence the RES-scores. By comparing likes with likes, we account for this within the BoD model. Secondly, the conditional efficiency model allows us to examine non-parametrically the direction (i.e., favorable or unfavorable to the RES-scores) and significance of the background characteristics. The impact of the background variables can be deduced by nonparametrically regressing the ratio of the conditional and unconditional RES-scores, $RES_c^m(y|z)/RES_c^m(y)$, on the background characteristics $z$. Extending the work of Daraio and Simar (2005, 2007a), which allowed for a visualization of the impact of background characteristics, De Witte and Kortelainen (2008) proposed to use non-parametric bootstrap based procedures Racine et al. (2006) in order to obtain statistical inference. The obtained results are the non-parametric equivalent to the standard $t$-tests.

## The data

We estimate and explain research performance for all 81 researchers at the department Business Administration of the Hogeschool Universiteit Brussel (HUB; a university in Belgium) for the period 2006–2008.[21] The data for this study were collected from three different sources: the official research evaluations, administrative records, and a questionnaire administered to the evaluated researchers.

The official research evaluations comprised the output of the individual researchers on nine output criteria. The selection of the nine criteria was performed by the faculty board where it took more than 2 years (with debates between researchers and policy makers) to come to a consensus on the preferred mix of output criteria which most faithfully reflect the policy priorities of the department. An overview of the nine output criteria and their maximal weights (as determined by the Faculty Board) is given in Table 1.[22]

We recognize that alternative selections of output criteria are possible. However, we also believe that it is ultimately the responsibility of the faculty board (in dialogue with the researchers) to define a selection of output criteria that most faithfully reflects the chosen objectives of the department. As presented in Table 2, the distribution of publications is heavily and negatively skewed. For example, 48 of the 81 researchers did not succeed in

[21] The majority of studies in the academic literature include researchers from different fields of research in their analysis. This may cause a bias in the results due to significant differences between research areas. In our analysis, the homogeneity of the set of observations (i.e., only researchers from the department Business Administration of HUB) guarantees that results are less biased (due to less heterogeneity in the areas of research in which researchers at HUB are active).

[22] It is important to note that the robust and conditional BoD-model can also be applied in faculty performance evaluations in other fields of academic research (e.g., natural sciences) as well as with other research criteria/data such as hard scientometric data and indicators (e.g., impact factors of journals, citations, etc.).

**Table 2** Summary statistics for the 73 evaluated researchers

| Output | Intern. journal[a] | Intern. book | Nation. book | Research report | National journal | Promoter project | National proceedings | National book | Discussion paper |
|---|---|---|---|---|---|---|---|---|---|
| Average | 1.272 | 0.173 | 0.346 | 0.593 | 2.765 | 0.457 | 0.086 | 0.346 | 1.728 |
| SD | 2.450 | 0.628 | 0.824 | 3.053 | 2.959 | 2.092 | 0.283 | 0.824 | 2.598 |
| Minimum | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| First quartile | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Median | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 |
| Third quartile | 2 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 2 |
| Maximum | 14 | 4 | 4 | 27 | 13 | 14 | 1 | 5 | 14 |

| Characteristic | Gender | Age | PhD | Retention | Research | Teaching | Affiliated |
|---|---|---|---|---|---|---|---|
| Average | 0.296 | 40.815 | 0.864 | 0.829 | 0.359 | 0.446 | 0.383 |
| SD | 0.459 | 9.010 | 0.345 | 0.229 | 0.162 | 0.204 | 0.489 |
| Minimum | 0 | 27 | 0 | 0.45 | 0.1 | 0.087 | 0 |
| First quartile | 0 | 34 | 1 | 0.5 | 0.25 | 0.28 | 0 |
| Median | 0 | 40 | 1 | 1 | 0.3 | 0.4275 | 0 |
| Third quartile | 1 | 46 | 1 | 1 | 0.5 | 0.56 | 1 |
| Maximum | 1 | 66 | 1 | 1 | 0.8 | 0.945 | 1 |

| Motivation | Q1[b] | Q2 | Q3 | Q4 | Q5 |
|---|---|---|---|---|---|
| Average | 4.229 | 4.329 | 3.057 | 3.771 | 2.629 |
| SD | 0.783 | 0.675 | 1.034 | 0.618 | 1.364 |
| Minimum | 1 | 3 | 1 | 2 | 1 |
| First quartile | 4 | 4 | 2 | 3 | 1.25 |
| Median | 4 | 4 | 3 | 4 | 2 |
| Third quartile | 5 | 5 | 4 | 4 | 4 |
| Maximum | 5 | 5 | 5 | 5 | 5 |

N = 73 observations (initial sample of 81 observations)

[a] The explanation of the research outlets is presented in Table 1

[b] The explanation of the motivational questions is presented in Table 5

publishing any article in an international journal from the Thompson Master List during the evaluated period. Moreover, 60 of the 118 (approximately 50%) papers published in the journals considered under the first output criteria 'International publication (Thompson Master List)' are the work of only eight researchers (approximately 10% of the research faculty). Similar remarks hold for the other research output criteria. In total, ten researchers did not succeed in providing any research output for the nine criteria. Numerous studies reported similar findings of heavily and negatively skewed distributions of research output (Ramsden 1994; Daniel and Fisch 1990). Lotka (1926) was the first to study this phenomenon. He found that the number of people producing $n$ papers is approximately proportional to $1/n^2$.

A second data source was the administrative records from the department of personnel administration. These employee records contain information on the researcher's age, gender, retention (ratio of the amount of time that a researcher is contracted for to the maximum amount of time), whether or not he/she obtained a doctoral degree (dummy variable with 1: yes and 0: no; including this variable is typical to the particular setting as HUB used to be a college with an exclusive focus on teaching, whereas recently, thanks to the Bologna reforms, the university is more and more research oriented), teaching load (percentage of time assigned to teaching activities), time for research (percentage of time assigned to research activities), and whether or not he or she is affiliated to another research department outside HUB (dummy variable with 1: yes and 0: no).

The third data source was a questionnaire administered to the 81 researchers. This questionnaire was developed based on a survey used in a previous study of Marsh and Hattie (2002, pp. 636–637). In particular, we asked the researchers to indicate their level of agreement on several statements on their research and teaching abilities, teaching and research satisfaction, personal goals, intrinsic and extrinsic rewards for teaching and research, beliefs about the relationship between teaching and research (e.g., the time conflict between teaching and research), and the departmental ethos for research and teaching. We further complemented this questionnaire with a number of statements on their research collaborations and their opinion on the impact of the situation at home on their research performance. We used a five-point likert scale where 5 represented "strongly agree" and 1 represented "strongly disagree". Exceptions are the two statements where the academics are asked to rate their ability as a teacher and researcher under ideal conditions (i.e., no limits on time, resources, etc.,). The five-point Likert scale for these two statements ranged from 1 "very poor" to 5 "very good". Further, we asked the researchers to indicate the number of persons with whom they have engaged in research collaborations within the past 12 months (proxy for research collaboration). The selection of questions that we further used in our analysis is listed in Table 5. Usable responses were obtained from 73 staff (from a total of 81 members), representing a total response rate of more than 90%. Extensive bivariate analyses point out that there is not a selection bias among the 8 missing observations. Specifically, 2 persons have a protracted illness, 2 persons retired recently, 1 person moved to another university and 3 persons refused to cooperate (by ideological reasons). The research output of the 8 missing observations is not significantly different from the research output of the other 73 observations. The final data set consists as such of 73 observations.

## Results

Before estimating the research performance of 73 researchers at HUB by the outlined conditional, robust BoD model, we present the RES-scores as they would be computed by

**Table 3** Estimates of research performance in different model specifications (n = 73 researchers)

| | Arithmetic average of performance criteria | Weighted average of performance criteria | Robust BoD of performance criteria | Robust and conditional BoD Model 1[a] | Robust and conditional BoD Model 2[a] |
|---|---|---|---|---|---|
| Average | 0.157 | 0.159 | 0.570 | 0.693 | 0.843 |
| SD | 0.180 | 0.191 | 0.520 | 0.351 | 0.323 |
| Minimum | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| First quartile | 0.039 | 0.029 | 0.130 | 0.399 | 0.873 |
| Median | 0.098 | 0.098 | 0.439 | 0.792 | 0.993 |
| Third quartile | 0.196 | 0.194 | 0.870 | 1.000 | 1.000 |
| Maximum | 1.000 | 1.000 | 2.458 | 1.006 | 1.090 |

[a] The included variables in Models 1 and 2 are presented in Table 4

the traditional methods in the literature (i.e., by an arithmetic or weighted average, and without accounting for the background). The resulting RES-scores, as presented in Table 3, indicate an on average low relative performance. As one single person obtained 51 output items, the research performance of the others seems rather bleak. We observe from Table 3 that some researchers obtain a RES-score of 0. These researchers did not publish any paper during the examined period and hence receive the lowest possible evaluation score. Even if the output items are weighted by the weights determined at the Faculty Board, as presented in the second column, the performance of most researchers seems rather poor (with an average RES-score of 0.159). According to this weighted model, 75% of the researchers could improve his/her weighted RES-score by approximately 81% if he/she would work as efficient as the most efficient researcher in the sample.

If the research performance of the faculty members would be evaluated by the use of a similar computation method (recall from a previous footnote that the 'most efficient' researcher published 27 very similar reports), the RES-score would not been taken seriously. Moreover, there would be huge resistance in using similar RES-scores as an incentive device (e.g., reducing teaching load, increasing wage).

By allowing for 'personalized' and 'optimal' weight restrictions, the BoD model is clearly more attractive to the individual researchers. To a certain extent (i.e., the weight bounds), researchers are given some leeway in their publication outlets. As such, the BoD model is less restrictive than the arithmetic or weighted average. Moreover, in its robust version, the BoD model accounts for outlying observations (e.g., the researcher with 27 similar publications) without losing information due to removing researchers from the data set. Summary statistics, as presented in Table 3, indicate that 75% of the researchers could increase their publication performance by at least 13% if they would publish as efficient as the best performing researchers (i.e., third quantile RES-value of 0.87). Similar as before, researchers who did not publish anything during the evaluation period obtain the lowest possible RES-score of 0.

The evaluated researchers may still feel a significant reluctance against the RES-scores if they do not account for the background of the researcher (particularly given the research and practical evidence that background characteristics can have a considerable influence on the opportunities to do research). Therefore, in Model 1, we allow for heterogeneity among researchers by using the conditional and robust BoD model. In a 'relative competition' (e.g., for personnel decisions), by comparing comparable researchers, the RES-scores can

**Table 4** Direction and impact (by $p$ value) of background characteristics on research performance (n = 73 researchers)

| | Model 1 | | Model 2 | |
|---|---|---|---|---|
| | Direction | $p$ value | Direction | $p$ value |
| Researcher characteristics | | | | |
| Gender (female = 1) | Favorable | 0.018** | Favorable | 0.000*** |
| Having a PhD (=1) | Favorable | 0.044** | Favorable | 0.032** |
| Guest researcher at KU Leuven | Favorable | 0.000*** | Favorable | 0.002*** |
| Age | Favorable | 0.906 | Favorable | 0.888 |
| Employment conditions | | | | |
| Retention | Favorable | 0.001*** | Favorable | 0.992 |
| Research time | Favorable | 0.035** | Unfavorable | 0.380 |
| Teaching time | Unfavorable | 0.366 | | |
| Motivation[a] | | | | |
| Q1: research gives satisfaction | | | Favorable | 0.004*** |
| Q2: time is constraint | | | Favorable | 0.038** |
| Q3: salary increase | | | Favorable | 0.944 |
| Q4: rating of ability | | | Favorable | 0.000*** |
| Q5: most collaboration within own department | | | Favorable | 0.422 |

***, **, and * Significance at 1, 5, and 10% level, respectively

[a] The explanation of the motivational questions is presented in Table 5

be considered as 'more fair'. Besides the employment conditions as retention, teaching load and research time, Model 1 accounts for some researcher background characteristics as gender, age, PhD and guest researcher at KU Leuven (i.e., HUB recently joined the KU Leuven and, as such, some researchers are affiliated with KU Leuven).[23] In other words, Model 1 controls for both truly exogenous factors (such as age, gender) or factors which are exogenous to the researcher as they are a university decision (e.g., hiring faculty without PhD, retention). Although this set of background variables is not exhaustive, it contains the variables that the faculty board at HUB (i.e., a mixture of policy makers and researchers) consider as appropriate (although one can have distinct views). For example, given the particular situation of HUB, the faculty board feels that having a PhD plays a role in explaining research performances as having the degree was not a requirement for faculty members at HUB (although recently, all appointments are only PhD's). A similar observation yields for affiliated researchers with KU Leuven, as the affiliation depends on the collaboration between two research groups (as such, if there is not a similar research group at KU Leuven, it is impossible to be affiliated). Also the age and gender is considered to play a role in research performance. Accounting for this set of background variables, the conditional RES estimates increase dramatically. A larger group of researchers (75%) becomes significant while the median researcher can improve its research performance by

[23] Note again that we do not attempt to investigate the causality of the variables. There might arise (serious) endogeneity among the variables (in that, e.g., better performing researchers obtain a higher retention). The conditional BoD methodology only attempts to correct the RES-scores for background characteristics which are often considered by researchers to be crucial to their performance (and, therefore, should be included to make a 'fair' comparison of researchers). As such, the significance level of the background variables, as presented in Table 4, indicates which of the background variables influence the RES-scores.

21%. Table 4 presents the direction (i.e., favorable or unfavorable to the RES-scores) and impact (i.e., if the impact is not significant, the RES-scores do not account for this variable) of the background characteristics.

As a first class of variables, consider the impact of the researcher characteristics. First, observe that female researchers, on average, have better research performances (as measured by RES-scores). This observation contrasts to previous parametric findings in the literature which indicated that men perform in general better than females (e.g., Ramsden 1994; Maske et al. 2003; Sax et al. 2002) or that there is no significant difference in research performances of men and females (e.g., Davis et al. 2001; Lee and Bozeman 2005). Secondly, note the favorable and significant impact of having a PhD on the RES-scores. Among others, Fox and Milbourne (1999) and Rodgers and Neri (2007) reported similar results. Thirdly, researchers who are affiliated at KUL (and, thus cooperate with researchers at KUL) realize higher RES-scores compared to their non-affiliated colleagues. This result confirms the findings of some previous parametric studies which indicate that research collaboration has a favorable influence on research performances (e.g., Ramos et al. 2007; Ramsden 1994; Godin and Gingras 2000). Finally, researcher age seems to have no significant impact on obtained RES-scores. This is in line with the findings of some previous parametric studies (e.g., Ramsden 1994; Gupta et al. 1999). However, as presented in Table 1, there are also other (parametric) studies which claimed that the relationship between the researcher's age and performance is on average negative (e.g., Sax et al. 2002; Levin and Stephan 1991).

As a second class of background variables, consider the influence of the employment conditions. Firstly, we observe a positive and significant influence of retention on measured RES-scores. Stated differently, academics who work full-time at HUB realize better research performers relative to their colleagues who only work part-time at HUB. This result confirms the finding of several parametric studies (e.g., Feldman 1983; Chen et al. 2006; Fairweather 2002). Further, we find that researchers who spent more time on research are more productive (i.e., higher RES-scores). This positive association is in line with the majority of the previous (parametric) studies in the literature. In fact, we found only two studies that claimed a non-significant relationship (i.e., Marsh and Hattie 2002 and Liddle et al. 1997). The (official) time spent on teaching is not significantly related to RES-scores. Although this contrasts to the findings of several studies and general believes (see Table 1), we found a few studies which observed similar outcomes (i.e., Hattie and Marsh 1996, 2002; Olsen and Simmons 1996).

To explore the impact of motivation on research performance, we estimate in Model 2 the RES-scores while accounting for the individual background characteristics and a set of motivational variables (as obtained from the questionnaires). From the larger questionnaire we performed among the evaluated researchers, we deduced 5 relevant questionnaire items which are described in Table 5.

Some summary statistics are provided in Table 2. Recall that a value of 5 denotes 'strongly agree', while a value of 1 denotes 'strongly disagree'. The results, as presented in Table 4, indicate that once accounted for motivation, employment conditions (i.e., retention, research time, and teaching time) do not longer play a role in the RES-scores (i.e., the bandwidth of the employment conditions becomes very large such that the impact is faded out). However, we do observe that the research characteristics 'gender', 'having a PhD' and 'guest researcher at KU Leuven' still have a considerable positive influence on research performances (as measured by RES-scores). Moreover, we find that the more satisfaction a researcher obtains from doing research, the better his/her research performances are. Several (parametric) studies denoted a similar positive association (Fox 1992;

**Table 5** Questionnaire items

| | |
|---|---|
| Q1 | Being involved in research gives me a great deal of satisfaction |
| Q2 | Time is a major constraint to improving my research productivity |
| Q3 | Having a salary increase related to my research performance would inspire me to become a better researcher |
| Q4 | Under ideal conditions (i.e., no limits on time, resources, etc.,), how do you rate your ability as a researcher? |
| Q5 | Most of my collaboration partners are members of my own department (HUB)? |

Chen et al. 2006; Sax et al. 2002). The researcher's perception as to whether time is a major constraint to improving his/her research productivity is positively and significantly related to RES-scores. This result is somewhat counterintuitive because it is usually thought that researchers, who experience a time constraint in the improvement of their research productivity, are less productive. Further, academics who rate their ability as a researcher (under ideal conditions) high, on average, perform better compared to colleagues who have a less positive view on their own research abilities. In other words, an optimistic self-image (with respect to one's own research abilities) contributes positively to the research performances. Vasil (1992, 1996) and Marsh and Hattie (2002) reported similar findings. Lastly, both having a salary increase and having most of its collaboration partners within the own institution (here, HUB) have non-significant influences on RES-scores.

## Conclusion

Given the increasing attention to research, universities and colleges are increasingly interested in evaluating the research performances of academics. Contrary to traditional single criterion measures, such as number of publications and citation counts, we suggested multi-criteria measures as they are more able to grasp the complex nature of research performances. In particular, we proposed a specially tailored version of the 'benefit-of-the-doubt' (BoD) model (which is rooted in the popular non-parametric Data Envelopment Analysis approach). The model is used to (1) integrate the performances of researchers on several research criteria into one overall Research Evaluation Score (RES-scores) while accounting for researcher characteristics and motivations as well as working conditions, and (2) non-parametrically analyze the impact (both in terms of direction and size) of these characteristics, motivations and working conditions on the RES-scores. In the context of constructing fair and robust RES-scores, this BoD approach has several advantages. First, for each individual researcher, weights for the output criteria are chosen such that the highest RES-score is realized. Secondly, the BoD model is flexible to incorporate the opinion of the faculty board and other stakeholders (including scientometricians) in the built-up of the RES-scores through a priori specified weight restrictions. Thirdly, the BoD model can be adjusted such that the resulting RES-scores are less sensitive to influences of outlying or extreme observations as well as potential measurement error in the data. Finally, the BoD model can be extended to account for several background influences (i.e., researcher characteristics and motivations as well as working conditions). Particularly with

an eye towards evaluations of research personnel, each of these advantages is essential for the credibility and acceptance of RES-scores.

To non-parametrically analyze the impact of research characteristics, motivations and employment conditions on RES-scores, we applied the bootstrap based *p*-values of De Witte and Kortelainen (2008). The results indicate that, on average, higher RES-scores are given to researchers who (a) are female, (b) have a PhD, (c) are affiliated with one or more other universities (here, mainly KU Leuven), (d) get more satisfaction out of doing research, (e) perceive that timing is a major constraint to improve their research, (f) rate their ability as a researcher higher. On the other hand, alternative examined background characteristics (i.e., researcher age, retention, research time, salary increase, and collaboration within own department) did not significantly influence measured RES-scores.

From point of view of the university management, this information is potentially useful. For instance, knowing that being affiliated to other universities can have a positive influence on research productivity, faculty boards may motivate researchers to collaborate with academics outside the own university. In addition, the board might consider stimulating researchers without a doctoral degree to attend a PhD program as this may enhance their research productivity in the future. Management can also attempt to improve the personnel policy and the working environment with an eye towards increasing research satisfaction and, thus research productivity. Finally, although sometimes claimed differently, (intrinsic) motivation is more important than salary increases.

Without doubt, the impact of research characteristics, motivations and employment conditions on research productivity will vary with the research area as well as other particular circumstances and conditions. Therefore, an interesting agenda for future research would be to apply the proposed BoD methodology in other evaluation settings to check for recurring patterns in the results. Similarly, it would be interesting to use the non-parametric BoD method to the data of previous parametric studies to compare the results. In case of different results, at first sight, the results of BoD could be preferred as no a priori assumptions are made. Another suggestion would be to extend the analysis with other potentially influential characteristics. Moreover, we believe that proposed method is well-suited to study the complex research-teaching nexus more profoundly. Finally, although not being a consideration of this paper, we stress the importance of studying the exact mechanisms by which aforementioned characteristics, motivations and conditions influence RES-scores more in detail.

**Acknowledgments** We are grateful to Emily Brounts, Laurens Cherchye, Wim Groot, Henriëtte Maassen van den Brink, Tom Van Puyenbroeck, Jessica Wery, seminar participants at the University of Amsterdam, and conference participants at the North American Productivity Workshop 2010 (Houston, USA) and 3rd workshop on efficiency and productivity analysis (Porto, Portugal) for valuable comments on a previous version of the paper.



## References

Allen, M. (1996). Research productivity and positive teaching evaluations: Examining the relationship through meta-analysis. *Journal of the Association for Communication, 2*, 77–96.

Archibugi, D., & Coco, A. (2004). A new indicator of technological capabilities for developed and developing countries (ArCo). *World Development, 32*(4), 629–654.

Astin, A. W. (1994). *What matters in college? Four critical years revisited*. San Francisco: Jossey-Bass Inc.
Avital, M., & Collopy, F. (2001). Assessing research performance: Implications for selection and motivation. *Sprouts: Working Papers on Information Environments, 1*(3), 40–61.
Babu, A. R., & Singh, Y. P. (1998). Determinants of research productivity. *Scientometrics, 43*(3), 309–329.
Badin, L., Daraio, C., & Simar, L. (2010). Optimal bandwidth selection for conditional efficiency measures: A data-driven approach. *European Journal of Operational Research, 201*(2), 633–640.
Beasly, J. E. (1995). Determining teaching and research efficiencies. *The Journal of the Operational Research Society, 46*(4), 441–452.
Becker, W. E., & Kennedy, P. E. (2005). Does teaching enhance research in economics. *Perspectives on Research and Teaching in Economics, 95*(2), 172–176.
Bellas, M. L., & Toutkoushian, R. K. (1999). Faculty time allocations and research productivity: Gender, race and family effects. *The Review of Higher Education, 22*, 367–390.
Bland, C. J. (1992). Characteristics of a productive research environment: Literature review. *Academic Medicine, 67*(6), 385–397.
Bland, C. J., & Berquist, W. H. (1997). The vitality of senior faculty members. Snow on the roof-fire in the furnace. ERIC Document Reproduction, service no. ED415733.
Bonaccorsi, A., Daraio, C., & Simar, L. (2006). Advanced indicators of productivity of universities: An application of robust nonparametric methods to Italian data. *Scientometrics, 66*(2), 389–410.
Bowen, H., Moesen, W., & Sleuwaegen, L. (2008). A composite index of the creative economy with application to regional best practices. *Review of Business and Economics, 53*(4), 375–397.
Braxton, J. M. (1996). Contrasting perspectives on the relationship between teaching and research. In *New directions for institutional research* (vol. 2, pp. 5–14). San Francisco: Jossey-Bass.
Brew, A. (2001). *The nature of research: Inquiry in academic contexts*. London: Routlegde/Falmer.
Broder, I. E. (1993). Professional achievements and gender differences among academic economists. *Economic Inquiry, 31*, 116–127.
Cazals, C., Florens, J. P., & Simar, L. (2002). Nonparametric frontier estimation: A robust approach. *Journal of Econometrics, 106*(1), 1–25.
Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*, 429–444.
Chen, Y., Gupta, A., & Hoshower, L. (2006). Factors that motivate business faculty to conduct research: An expectancy theory analysis. *Journal of Education for Business, 81*, 179–189.
Cherchye, L., & Kuosmanen, T. (2006). Benchmarking sustainable development: A synthetic meta-index approach, Chapter 7. In M. McGillivray & M. Clarke (Eds.), *Perspectives on human development*. Tokyo: United Nations University Press.
Cherchye, L., Lovell, C. A. K., Moesen, W., & Van Puyenbroeck, T. (2007a). One market, one number? A composite indicator assessment of EU internal market dynamics. *European Economic Review, 51*(3), 749–779.
Cherchye, L., Moesen, W., Rogge, N., & Van Puyenbroeck, T. (2007b). An introduction to 'benefit of the doubt' composite indicators. *Social Indicators Research, 82*, 111–145.
Cherchye, L., Moesen, W., Rogge, N., Van Puyenbroeck, T., Saisana, M., Saltelli, A., et al. (2008). Creating composite indicators with DEA and robustness analysis: The case of the technology achievement index. *Journal of the Operational Research Society, 59*, 239–251.
Cherchye, L., & Vanden Abeele, P. (2005). On research efficiency: A micro-analysis of Dutch university research in economics and business management. *Research Policy, 34*, 495–516.
Claro, J., & Costa, C. A. V. (2010). A made-to-measure indicator for cross-disciplinary bibliometric ranking of researchers performance. *Scientometrics*. doi:10.1007/s11192-010-0241-5.
Cook, W. D. (2004). Qualitive data in DEA. In W. W. Cooper, L. Seiford, & J. Zhu (Eds.), *Handbook on data envelopment analysis* (pp. 75–97). Dordrecht: Kluwer Academic Publishers.
Daniel, H.-D., & Fisch, R. (1990). Research performance evaluation in the German university sector. *Scientometrics, 19*(5–6), 349–361.
Daouia, A., & Simar, L. (2007). Nonparametric efficiency analysis: A multivariate conditional quantile approach. *Journal of Econometrics, 140*(2), 375–400.
Daraio, C., & Simar, L. (2005). Introducing environmental variables in nonparametric frontier models: A probabilistic approach. *Journal of Productivity Analysis, 24*(1), 93–121.
Daraio, C., & Simar, L. (2007a). *Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications*. Series: Studies in Productivity and Efficiency. Dordrecht: Springer.
Daraio, C., & Simar, L. (2007b). Conditional nonparametric frontier models for convex and nonconvex technologies: A unifying approach. *Journal of Productivity Analysis, 28*, 13–32.
Dasgupta, P., & David, P. A. (1994). Toward a new economics of science. *Research Policy, 23*(5), 487–521.

Davis, J. C., Huston, J. H., & Patterson, D. M. (2001). The scholarly output of economists: A description of publishing patterns. *Atlantic Economic Journal, 29*(3), 341–349.
Davis, J. C., & Patterson, D. M. (2001). Determinants of variations in journal publication rates of economists. *American Economist, 45*, 86–91.
De Witte, K., & Kortelainen, M. (2008). Blaming the exogenous environment? Conditional efficiency estimation with continuous and discrete environmental variables, CES Discussion Paper Series DPS 08.33, MPRA Paper 14034.
De Witte, K., & Rogge, N. (2009). Accounting for exogenous influences in a benevolent performance evaluation of teachers, Brussel. HUB Research Papers 2009/15.
Despotis, D. K. (2005). A reassessment of the human development index via data envelopment analysis. *Journal of the Operational Research Society, 56*(8), 969–980.
Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education, 11*(1), 37–46.
European Commission. (2004). The EU Economy Review 2004, European Economy, Nr. 6 .Luxembourg: Office for Official Publications of the EC.
Fabel, O. (2008). Research productivity in business economics: An investigation of Austrian, German and Swiss universities. *German Economic Review, 9*(4), 506–531.
Fairweather, J. S. (1993). Faculty rewards reconsidered: The nature of tradeoffs. *Change, 25*, 44–47.
Fairweather, J. S. (2002). The mythologies of faculty productivity: Implication for institutional policy and decision making. *The Journal of Higher Education, 73*(1), 26–48.
Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society, Series A*, *CXX*, Part 3, 253–290.
Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education, 18*, 3–124.
Feldman, K. A. (1987). Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: A review and exploration. *Research in Higher Education, 26*, 227–298.
Foster, J., & Sen, A. (1997). *On economic inequality*, 2nd expanded edn. Oxford: Clarendon Press.
Fox, M. F. (1992). Research, teaching, and publication productivity: Mutuality versus competition in academia. *Sociology of Education, 65*, 293–305.
Fox, M. F., & Faver, C. A. (1985). Men, women, and publication productivity: Patterns among social work academics. *The Sociological Quarterly, 26*(4), 537–549.
Fox, M. F., & Milbourne, R. (1999). What determines research output of academic economists? *Economic Record, 75*, 256–267.
Godin, B., & Gingras, Y. (2000). Impact of collaborative research on academic science. *Science and Public Policy, 27*(1), 65–73.
Gupta, B. M., Kumar, S., & Aggarwal, B. S. (1999). A comparison of productivity of male and female scientists of CSIR. *Scientometrics, 45*(2), 269–289.
Harris, G., & Kaine, G. (1994). The determinants of research performance: A study of Australian university economists. *Higher Education, 27*, 191–201.
Hattie, J., & Marsh, H. W. (1996). The relationship between research and teaching: A meta-analysis. *Review of Educational Research, 66*, 507–542.
Hattie, J., & Marsh, H. W. (2002). The relation between research productivity and teaching effectiveness. *Journal of Higher Education, 73*(5), 603–641.
Jacobsen, R. L. (1992). Colleges face new pressure to increase faculty productivity. *Chronicle of Higher Education, 38*(32), 16–18.
Jenkins, A., Blackman, T., Lindsay, R., & Patton-Saltzberg, R. (1998). Teaching and research: Student perspectives and policy implications. *Studies in Higher Education, 23*, 127–141.
Jeong, S., Park, B., & Simar, L. (2010). Nonparametric conditional efficiency measures: Asymptotic properties. *Annals of Operations Research, 173*(1), 105–122.
Johnes, G., & Johnes, J. (1993). Measuring the research performance of UK university departments: An application of data envelopment analysis. *Oxford Economic Papers, 45*, 332–347.
Kremer, J. (1990). Construct validity of multiple measures in teaching, research and services and reliability of peer ratings. *Journal of Educational Psychology, 82*, 213–218.
Kuosmanen, T. (2002). Modeling blank entries in Data Envelopment Analysis. EconWPA Working Paper at WUSTL, No. 0210001.
Kyvik, S. (1990). Age and scientific productivity. Differences between fields of learning. *Higher Education, 19*, 37–55.
Landry, R., Traore, N., & Godin, B. (1996). An econometric analysis of the effect of collaboration on academic research productivity. *Higher Education, 32*(3), 283–301.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science, 35*(5), 673–702.
Levin, S. G. S., & Stephan, P. (1991). Research productivity over the life cycle: Evidence for academic scientists. *American Economic Review, 81*(1), 114–132.
Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton: Princeton University Press.
Liddle, B., Westergren, A., & Duke, D. (1997). Time allocation and research productivity among counseling faculty. *Psychological Reports, 80*, 339–344.
Long, J. S. (1987). Problems and prospects for research and sex differences in the scientific career. In L. S. Dix (Ed.), *Women: Their underrepresentation and career differentials in science and engineering* (pp. 157–169). Washington: National Academy Press.
Long, J. S., Allison, P. D., & McGinnis, R. (1993). Rank advancement in academic careers: Sex differences and the effects of productivity. *American Sociological Review, 58*(5), 703–722.
Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*(12), 317–324.
Lovell, C. A. K., Pastor, J. T., & Turner, J. A. (1995). Measuring macroeconomic performance in the OECD: A comparison of European and Non-European Countries. *European Journal of Operational Research, 87*, 507–518.
Marsh, H. W. (1984). Students' evaluations of teaching: Dimensionality, reliability, validity, potential biases and utility. *Journal of Educational Psychology, 76*, 707–754.
Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253–388.
Marsh, H. W., & Hattie, J. (2002). The relationship between research productivity and teaching effectiveness: Complementarity, antagonistic, or independent constructs? *The Journal of Higher Education, 73*(5), 603–641.
Martin, B. R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics, 36*(3), 343–362.
Maske, K. L., Durden, G. C., & Gaynor, P. (2003). Determinants of scholarly productivity among male and female economists. *Economic Inquiry, 41*, 555–564.
Massy, W. F., & Wilger, A. K. (1995). Improving productivity. *Change, 27*(4), 10–20.
McCain, K., & Turner, K. (1989). Citation context analysis and aging patterns of journal articles in molecular genetics. *Scientometrics, 17*(1/2), 127–163.
Melyn, W., & Moesen, W. (1991). Towards a synthetic indicator of macroeconomic performance: Unequal weighting when limited information is available, Public Economics Research Paper 17. Leuven: CES, KULeuven.
Mercer, B. E., & Wanderer, J. J. (1970). *The study of society*. Belmont: California Wardsworth Publishing Co.
Merton, R. K. (1968). *Social theory and social structure*. New York: The Free Press.
Nederhof, A. J., Meijer, R. F., Moed, H. F., & Vanraan, A. F. J. (1993). Research performance indicators for university departments: A study of an agricultural university. *Scientometrics, 27*(2), 157–178.
Neumann, R. (1992). Perceptions of the teaching-research nexus: A framework for analysis. *Higher Education, 23*, 159–171.
OECD, European Commission, Joint Research Centre. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. In Nardo, M., Saisana, M., Saltelli, A., & Tarantola, S. (EC/JRC), Hoffman, A. & Giovannini, E. (OECD), Paris: OECD publication code: 302008251E1.
Olsen, D., & Simmons, A. J. M. (1996). The research versus teaching debate: Untangling the relationships. *New Directions for Institutional Research, 90*, 31–39.
Racine, J. S., Hart, J., & Li, Q. (2006). Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews, 25*(4), 523–544.
Ramos, R., Royuela, V., & Suriñach, J. (2007). An analysis of the determinants in economics and business publications by Spanish universities between 1994 and 2004. *Scientometrics, 71*(1), 117–144.
Ramsden, P. (1994). Describing and explaining research productivity. *Higher Education, 28*, 207–226.
Ramsden, P., & Moses, I. (1992). Association between research and teaching in Australian higher education. *Higher Education, 23*, 273–295.
Rodgers, J. R., & Neri, F. (2007). Research productivity of Australian academic economists: Human-capital and fixed effects. *Australian Economic Papers, 46*(1), 67–87.
Rogge, N. (2009a). Granting teachers the 'benefit-of-the-doubt' in performance evaluations, Brussels. HUB Research Papers 2009/17.
Rogge, N. (2009b). Robust benevolent evaluations of teacher performance. In *Conference proceedings: SIS 2009—colloquium 'simulation in industry and services 2009'*, Brussels. 4 December 2009.

Ruiz, C. F., Bonilla, R., Chavarro, D., Orozco, L. A., Zarama, R., & Polanco, X. (2010). Efficiency measurement of research groups using data envelopment analysis and bayesian networks. *Scientometrics, 83*(3), 711–721.

Rushton, J. P., Murray, H. G., & Paunonen, S. V. (1983). Personality, research creativity, and teaching effectiveness in university professors. *Scientometrics, 5*, 93–116.

Sax, L. J., Hagedorn, L. S., Arredondo, M., & Dicrisi, F. A. (2002). Faculty research productivity: Exploring the role of gender and family-related factors. *Research in Higher Education, 43*(4), 423–446.

Seldin, P. (1985). *Changing practices in faculty evaluation*. San Francisco: Jossey-Bass.

Stack, S. (2003). Research productivity and student evaluation of teaching in social science classes: A research note. *Research in Higher Education, 44*(5), 539–556.

Van Leeuwen, T. H., Visser, M. S., Moed, H. F., Nederhof, T. J., & Van Raan, A. F. J. (2003). The Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics, 57*(2), 257–280.

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics, 36*(3), 397–420.

Vasil, L. (1992). Self-efficacy expectation and causal attributions for achievement among male and female university faculty. *Journal of Vocational Behavior, 41*(3), 259–269.

Vasil, L. (1996). Social process skills and career achievement among male and female academics. *Journal of Higher Education, 67*(1), 103–114.

Vinkler, P. (1998). General performance indexes calculated for research institutes of the Hungarian academy of sciences based on scientometric indicators. *Scientometrics, 41*(1–2), 185–200.

Vinkler, P. (2006). Composite scientometric indicators for evaluating publications of research institutes. *Scientometrics, 68*(3), 629–642.

Volkwein, J. F., & Carbone, D. A. (1994). The impact of departmental research and teaching climates on undergraduate growth and satisfaction. *Journal of Higher Education, 65*, 147–167.

Zamarripa, E. J. (1995). Evaluating research productivity. *SRA Journal, 26*(3–4), 17–27.

Zhou, P., Ang, B. W., & Poh, K. L. (2007). A mathematical programming approach to constructing composite indicators. *Ecological Economics, 62*, 291–297.