European Journal of Operational Research 196 (2009) 847–855

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Invited Tutorial

Comparative evaluation of performance of national R&D programs with heterogeneous objectives: A DEA approach

Hakyeon Lee [a], Yongtae Park [a,*], Hoogon Choi [b]

[a] Department of Industrial Engineering, School of Engineering, Seoul National University, San 56-1, Shillim-dong, Kwanak-gu, Seoul 151-742, Republic of Korea
[b] Department of Systems Management Engineering, School of Engineering, Sungkyunkwan University, 300 Cheoncheon-dong, Jangan-gu, Suwon, Gyeonggi-do 440-746, Republic of Korea

ARTICLE INFO

Article history:
Received 3 September 2007
Accepted 19 June 2008
Available online 26 June 2008

Keywords:
Data envelopment analysis
R&D evaluation
R&D performance
National R&D program
Analytic hierarchy process

ABSTRACT

The strategic importance of performance evaluation of national R&D programs is highlighted as the resource allocation draws more attention in R&D policy agenda. Due to the heterogeneity of national R&D programs' objectives, however, it is intractably difficult to relatively evaluate multiple programs and, consequently, few studies have been conducted on the performance comparison of the R&D programs. This study measures and compares the performance of national R&D programs using data envelopment analysis (DEA). Since DEA allows each DMU to choose the optimal weights of inputs and outputs which maximize its efficiency, it can mirror R&D programs' unique characteristics by assigning relatively high weights to the variables in which each program has strength. Every project in every R&D program is evaluated together based on the DEA model for comparison of efficiency among different systems. Kruskal–Wallis test with a post hoc Mann–Whitney *U* test is then run to compare performance of R&D programs. Two alternative approaches to incorporating the importance of variables, the AR model and output integration, are also introduced. The results are expected to provide policy implications for effectively formulating and implementing national R&D programs.

1. Introduction

As R&D has been considered as a driving force for national competitive advantage, many countries have been raising R&D investments through various national R&D programs (Lee et al., 1996). Since R&D investment is one of the most decisive elements in promoting scientific and technological progress (Wang and Huang, 2007), the effective use of the limited R&D resources can be regarded as a prerequisite for benefiting from formulation and implementation of national R&D programs. Thus, performance evaluations of R&D programs need to be made so that the limited resources are allocated to promising R&D programs and poor R&D programs can be improved or terminated.

Although a number of studies have been conducted to measure R&D performance at various levels, few attempts have been made at the national program-level. This is due to the heterogeneity of R&D programs in terms of policy purpose. Since each R&D program has its own primary objective such as publishing academic papers for basic research, issuing patents and developing prototypes for applied research, and providing funds with researchers for R&D human resource development, it is intractably difficult to relatively compare the performance of various national R&D programs at the same time and in the same context.

Two conventional approaches to assessing R&D performance, peer review and bibliometric method do not work well for the relative evaluation of heterogeneous R&D programs. The peer review method, which is based on perceptions of well-informed experts about various quality dimensions of R&D, is inherently subjective and likely to be biased depending on interests, experience, and knowledge of the evaluators (Nederhof and van Raan, 1987; Brinn et al., 1996). The bibliometric method is considered relatively objective, but the results highly depend on the measurement method (Nederhof and van Raan, 1993). Regardless of whether simple count of publications is employed or the number of papers is adjusted based on citations or impact scores, the conventional bibliometric method has a limitation in that it only deals with a single type of R&D outputs, namely, publications. Although it is suitable to evaluation of basic research program or university research, other important outputs of national R&D activities such as human resources are ignored. A problem still occurs even when considering multiple outputs of R&D activities. To incorporate several output variables and produce a single measure for performance comparison, the relative importance of variables needs to be determined and fixed. However, it does not make sense that the same set of weights is applied to evaluations of R&D programs with different objectives.

* Corresponding author. Tel.: +82 2 880 8358; fax: +82 2 878 3511.
E-mail address: parkyt@cybernet.snu.ac.kr (Y. Park).

0377-2217/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2008.06.016

The tenet of this paper is data envelopment analysis (DEA) can overcome these limitations. DEA is a linear programming model for measuring the relative efficiency of decision making units (DMUs) with multiple inputs and outputs (Cooper et al., 2000). Since it can not only handle multiple outputs, but also allow each DMU to choose the optimal weights of inputs and outputs which maximize its efficiency (Cherchye et al., 2007), it is capable of mirroring R&D programs' unique characteristics by assigning high weights to the variables in which each project has strength. This study measures and compares the performance of six national R&D programs in Korea using the DEA efficiency. Basically, the performance of a R&D program is measured based on the performance of projects belonging to the program. DEA is carried out for the whole set of R&D projects in the six R&D program. Kruskal–Wallis test with a post hoc Mann–Whitney U test is then run to compare performance of R&D programs.

The remainder of this paper is organized as follows. Section 2 reviews the DEA models used in this study and applications of DEA to measuring R&D performance. The data for the case study are briefly introduced in Section 3. Section 4 deals with the details of analysis and results. The paper ends with conclusions in Section 5.

2. DEA

2.1. DEA models

DEA is a non-parametric approach that does not require any assumptions about the functional form of a production function and *a priori* information on importance of inputs and outputs. The relative efficiency of a DMU is measured by estimating the ratio of weighted outputs to weighted inputs and comparing it with other DMUs. DEA allows each DMU to choose the weights of inputs and outputs which maximize its efficiency. The DMUs that achieve 100% efficiency are considered efficient while the other DMUs with efficiency scores below 100% are inefficient.

The first DEA model proposed by Charnes et al. (1978) is the CCR model that assumes that production exhibits constant returns to scale. Banker et al. (1984) extended it to the BCC model for the case of variable returns to scale. When it comes to R&D returns to scale, findings from previous studies are somewhat mixed (Graves and Langowitz, 1996). It was found that R&D activity can exhibit increasing or decreasing returns to scale as well as constant returns to scale (Bound et al., 1984; Scherer, 1983); thus, the BCC model is employed in this study. DEA models are also distinguished by the objective of a model: maximize outputs (output-oriented) or minimize inputs (input-oriented). It is implicitly assumed that the objective of R&D lies in increasing outputs rather than decreasing inputs. Therefore, this study adopts the output-oriented model.

The output-oriented BCC model is formulated as

$$\begin{aligned} \max\quad & \eta \\ \text{s.t.}\quad & \mathbf{X}\boldsymbol{\lambda} \leqslant \mathbf{x}_o, \\ & \eta\mathbf{y}_o - \mathbf{Y}\boldsymbol{\lambda} \leqslant 0, \\ & \mathbf{e}\boldsymbol{\lambda} = 1, \\ & \boldsymbol{\lambda} \geqslant 0, \end{aligned} \tag{1}$$

where $\mathbf{X}$ is the matrix of input vectors, $\mathbf{Y}$ is the matrix of output vectors, $(\mathbf{x}_o, \mathbf{y}_o)$ is the DMU being measured, η is the reverse of the efficiency score, λ is the vector of intensity variables. The only difference between the CCR and BCC model is the presence of the convexity condition, $\mathbf{e}\boldsymbol{\lambda} = 1$. The dual form is

$$\begin{aligned} \min\quad & z = \mathbf{v}\mathbf{x}_o - v_o \\ \text{s.t.}\quad & \mathbf{u}\mathbf{y}_o = 1, \\ & \mathbf{v}\mathbf{X} - \mathbf{u}\mathbf{Y} - v_o\mathbf{e} \geqslant 0, \\ & \mathbf{v} \geqslant 0,\ \mathbf{u} \geqslant 0,\ v_o,\ \text{free in sign}, \end{aligned} \tag{2}$$

where $\mathbf{v}$ and $\mathbf{u}$ are the vectors of the weights given to inputs and outputs, respectively, v_o is the scalar associated with $\mathbf{e}\boldsymbol{\lambda} = 1$ in the primal form. The decision variables of the problem are the values of the weights ($\mathbf{v}$ and $\mathbf{u}$). To put it differently, the weights are chosen in a manner that assigns a best set of weights to each DMU where "best" means the resulting efficiency score is maximized under the given data. Thus, the weights differ across DMUs. Such a flexible and endogenous weighting system of DEA is called the "benefit of the doubt" approach (Cherchye et al., 2007). One thing that should be noted is the optimal weights at a given point, particularly an extreme efficient point, may not be unique (Rosen et al., 1998). This non-uniqueness problem may cause problems with interpretation.

If prior knowledge or accepted views exist, however, the weight flexibility in DEA leads to produce unrealistic efficiency scores (Allen et al., 1997). When this is the case, restrictions need to be placed on weights in DEA (Dyson and Thanassoulis, 1988). The use of the weights restriction makes it possible to mirror preference in a real world. Among a large diversity of weights restriction methods, the most common one is the assurance regions (AR) model proposed by Thompson et al. (1990). The AR of type 1 is to impose restrictions on the upper bound (U_{ij}) and lower bound (L_{ij}) of a ratio of the weights of two variables (u_i/u_j) as the following:

$$L_{ij} \leqslant \frac{u_i}{u_j} \leqslant U_{ij}. \tag{3}$$

This study also employs the AR model to reflect the relative importance of the same types of output variables.

One general assumption of DEA models is convexity. It means the production possibility set P is convex so that if two activities, $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$, belong to P, then every point on the line segment between these two points belongs to P. However, when DMUs belonging to different systems are considered and compared together, the convexity assumption is not valid since different systems may use different activities which cannot be compared. In such a situation, the convexity assumption only holds within the same system but does not hold between different systems (Cooper et al., 2000). In order to formulate the corresponding model, input and output vectors are divided for each system. That is, in the case of two systems, A and B, input $\mathbf{x}$ is divided into $\mathbf{x}_A$ and $\mathbf{x}_B$, and output $\mathbf{y}$ into $\mathbf{y}_A$ and $\mathbf{y}_B$. Also, the binary variables, z_A and z_B, are added as a dummy which can only assume the values 0 or 1. The output-oriented BCC model for comparison of efficiency between two systems is written as the following mixed integer linear programming problem:

$$\begin{aligned} \max\quad & \eta \\ \text{s.t.}\quad & \mathbf{X}_A\boldsymbol{\lambda}_A + \mathbf{X}_B\boldsymbol{\lambda}_B \leqslant \mathbf{x}_o, \\ & \eta\mathbf{y}_o - (\mathbf{Y}_A\boldsymbol{\lambda}_A + \mathbf{Y}_B\boldsymbol{\lambda}_B) \leqslant 0, \\ & \mathbf{e}\boldsymbol{\lambda}_A = z_A, \\ & \mathbf{e}\boldsymbol{\lambda}_B = z_B, \\ & z_A + z_B = 1, \\ & \boldsymbol{\lambda}_A \geqslant 0, \quad \boldsymbol{\lambda}_B \geqslant 0, \\ & z_A, z_B = 0 \text{ or } 1. \end{aligned} \tag{4}$$

This model ensures not only the evaluation of the efficiency of each DMU but also the comparison of the two systems. Although the above model deals with two systems, it can be generalized to more number of systems. This study employs the different system model by extending it to six systems.

2.2. DEA for measuring R&D performance

DEA has some attractive features in measuring R&D performance (Wang and Huang, 2007). First, DEA can be utilized even in the situation where *a priori* information on preference about variables does not exist. This is exactly the context of R&D performance evaluation in which there is no universally agreed view on importance of R&D inputs and outputs. As mentioned before, DEA solves this problem by automatically deriving the endogenous weights that represent a relative value system for each DMU (Allen et al., 1997). Such a data-oriented weighting method is considered justifiable and useful in the construction of composite indicators under the context of uncertainty about, and lack of consensus on an appropriate weighting scheme (Cherchye et al., 2007). The "benefit of the doubt approach" of DEA plays an even more crucial role for program-level evaluation than project-level evaluation. When R&D projects of the same program are evaluated, there may be an accepted view of preference on R&D outputs according to the main objective of the R&D program. Due to the heterogeneity of ultimate goals of various R&D programs, on the other hand, the relative importance of R&D outputs differ across programs. A published paper in an international journal could be more or less important than a granted patent or a student graduated with doctoral degree. The complete flexibility in the selection of weights in DEA not only makes it possible to compare performance of R&D programs with different objectives, but also take away an opportunity for excuse of inefficient R&D program managers because inefficient R&D programs are under-performing even by putting their own weights solely on "self esteem". Second, DEA is useful for the case in which the function of relationships between inputs and outputs cannot be defined. As mentioned before, since DEA does not require any assumptions about the functional form of a production function, it fits R&D activities whose production functions have not been specified. Third, DEA can handle multiple inputs and outputs which cannot be dealt with in the standard parametric methods. It enables various types of inputs and outputs of R&D activities to be considered in performance evaluation.

These advantages have led DEA to be effectively used for R&D performance evaluation at such various levels as nations (Wang and Huang, 2007; Kocher et al., 2006; Lee and Park, 2005), universities (Feng et al., 2004; Korhonen et al., 2001), academic disciplines (Garg et al., 2005), projects (Eilat et al., 2006; Swink et al., 2006), and project teams (Paradi et al., 2002). Nevertheless, the program-level application remains a void in the literature.

3. Data

3.1. R&D programs

This study deals with six national R&D programs supported by a government foundation located at Daejeon in Korea. The foundation aims to promote the creative potential of Korea's science and technology by formulating and supporting a variety of R&D programs. The annual budget of the foundation in 2007 exceeded USD 1 billion. The R&D programs supported by the foundation can be categorized into four major programs, and the selected six programs are sub-programs in one of the major programs. Though they belong to the same major program, the six programs have different objectives and different number of R&D projects, as shown in Table 1. The foundation has been evaluating individual projects based on the peer review method as well as quantitative outcomes, but it has never conducted performance comparison among the programs. In order for the limited budget to be effectively used, however, managers in the foundation feel it necessary to know which programs performed better or worse.

The primary goal of Program A and E is to develop outstanding human resources for basic science and bioengineering, respectively. Program C focuses on basic scientific research while Program D and F are aimed at developing national competitiveness

Table 1
Descriptions of R&D programs

Program	Number of projects	Objective
A	436	Supporting researchers to work in creative research activities individually or collaboratively in the fields of basic sciences emphasizing interdisciplinary research, and encouraging to develop outstanding human resources for research
B	46	Increasing competence in creative innovation, securing national core technologies which will enable to create new businesses
C	21	Supporting creative basic research to ensure exploration of basic theory and knowledge that may lead to outstanding papers and the development of advanced technologies
D	28	Supporting mainly in engineering research with the potential for industrial advancement, while encouraging interdisciplinary collaborations between industry and academia and enhancing the international competitiveness of national industries
E	13	Supporting large-scale, long-term R&D activities, the outputs of which are to be used both in bioengineering and clinical medicine, and playing a key role in cultivating human resources in basic medical sciences
F	4	Supporting research centers that are considered capable of creating world-class knowledge and competitiveness in core science and technology fields that require strategic incubation for national interests

through applied research. What is at the core of Program B is to create new business with innovative technologies.

All projects in Program A started in 2002 or in 2003 and ended in 2005. The projects in the other five programs were initiated between 1998 and 2002. Since some of those projects are still in progress in 2007, only the projects finished in 2005 were included in the data set. Although the duration and the time of start and end somewhat differ across projects, it is assumed that there is no effect of differences in project duration on the performance of projects since the amount of funds is basically determined in proportional to duration.

3.2. Variables

Two inputs and ten outputs were selected for this study, as shown in Table 2. The amount of funds given to a project and the number of researchers on a project were selected as a proxy for labor input and capital input, respectively. The amount of funds is the rest excluding the salary paid to researchers from total funds.

The four variables related with papers were included to the set of output variables. The academic papers published in journals have been considered as the major output of research and widely used to evaluate performance of researchers (OECD, 2001). Patents were also selected as outputs since patents have most frequently used as direct output of R&D activities (Zhang et al., 2003). The variables about patents are also divided into the four types. Finally, the output includes the perspective of human resource development of R&D with two variables, graduate students with master's degree and doctoral degree. Contrary to R&D projects of private firms, human resource development is one of the main objectives of governmental R&D programs (Garg et al., 2005).

The data were obtained from the foundations' database and arranged for analysis. The outputs produced between termination of projects and the end of 2007 were also counted since it usually takes a long time for R&D outputs to be realized (Adams and Griliches, 2000). The descriptive statistics of the data are given in Appendix A.

Table 2
Variables for DEA

Variables	Description
Input	
(I1) Funds	Total amount of funds given to a project
(I2) Researchers	Number of Ph.D. researchers on a project
Output	
(O1) Domestic SCI papers	Number of domestic scientific and technical articles published or accepted in journals listed on SCI (science citation index)
(O2) Domestic non-SCI papers	Number of domestic scientific and technical articles published or accepted in journals not listed on SCI
(O3) International SCI papers	Number of international scientific and technical articles published or accepted in journals listed on SCI
(O4) International non-SCI papers	Number of international scientific and technical articles published or accepted in journals not listed on SCI
(O5) Domestic applied patents	Number of patents applied in domestic patent offices
(O6) Domestic granted patents	Number of patents registered in domestic patent offices
(O7) Foreign applied patents	Number of patents applied in foreign patent offices
(O8) Foreign granted patents	Number of patents registered in foreign patent offices
(O9) Master's degree students	Number of students graduated with master's degree
(O10) Doctoral degree students	Number of students graduated with doctoral degree

4. Analysis and results

4.1. Measuring performance of R&D programs

This study aims to measure and compare the performance of national R&D programs with heterogeneous objectives using DEA. Although comparison is made at the program-level, a DMU in this study is a project, not a program. The reason for this is two-fold. If we consider R&D programs as DMUs, the data for each project have to be aggregated into single values for each program. The results then highly depend on how to aggregate data, and, regardless of the method, information loss occurs. The bigger problem is weak discriminatory power. When the number of DMUs is relatively small compared with the number of inputs and outputs, a large portion of the DMUs will be identified as efficient so the efficiency discrimination is lost (Cooper et al., 2000). The rule of thumb is that the number of DMUs should be at least three times larger than the combined number of inputs and outputs (Banker et al., 1989) or more than the product of the numbers of inputs and outputs (Boussofiane et al., 1991). In most situations, the number of R&D programs does not comply with the above rules since it is usually small compared with the number of inputs and outputs, as is in the case of this study. Then, most of the programs are likely to be considered equally efficient.

Thus, the reasonable alternative is to treat a R&D project as a DMU, not a R&D program. The project-level evaluation is conducted, and the performance of a program is then measured based on the performance of projects belonging to the program. Two options exist on the project-level evaluation. The first one is to run DEA for each R&D program. Only the projects belonging to the same program are evaluated together by DEA. Then, a measure or an indicator can be employed for performance comparison among programs. This kind of method for comparison has often been used when using traditional performance measures, but it does not make any sense in the case of the DEA efficiency since DEA measures the relative efficiency among DMUs included, not the absolute one. See the actual example in Table 3 summarizing the results of conducting DEA with the output-oriented BCC model for each program. In terms of both the average efficiency score and

Table 3
Results of independent evaluation of individual programs

Programs	Number of projects	Average efficiency score	Number of efficient projects (percentage)
A	436	0.4435	30 (6.88%)
B	46	0.6971	11 (23.91%)
C	21	0.8587	13 (61.90%)
D	28	0.9268	21 (75.00%)
E	13	0.9862	12 (92.31%)
F	4	1	4 (100%)

the portion of efficient projects, Program A has the lowest (0.4435, 6.88%) while Program F ranks the first (1, 100%). However, what makes differences is not the actual performance, but the number of projects included in each program. It is illustrated that the less the number of projects included, the higher both the average efficiency score and the percentage of efficient projects. The extreme case is found in Program F in which all of the four projects are evaluated as efficient. Even if the number of DMUs of programs is the same, the comparison cannot be made because projects are not evaluated at the same time.

To compare the performance among R&D programs, therefore, every project in every program has to be evaluated together. However, the six R&D programs are not directly comparable since their primary objectives are different. In other words, the basic BCC model cannot be used since the convexity assumption does not hold among the six programs. Thus, we employed the model for comparison of efficiency between different systems as the basic model for the analysis. An extension of (4) to six systems was applied to the whole set of 548 R&D projects of the six R&D programs. As a result, 72 projects (13.14%) were found to be efficient and the average efficiency score is 0.5146. Table 4 summarized the average efficiency scores of the six programs. Program E has the highest average score while Program F ranks the lowest. Since all the projects are evaluated together, it now makes sense to compare the performance of R&D programs based on the DEA results.

However, simple comparison based on average efficiency scores does not have statistical validity since theoretical distribution of the efficiency scores in DEA is unknown. Thus, Kruskal–Wallis test was run to compare performance of R&D programs. It is a non-parametric method that compares between the medians of two or more samples to determine if the samples have come from different populations. A post hoc Mann–Whitney U test was also conducted for paired comparisons. The results are also shown in Table 4. It was found that there exist statistically significant differences among efficiency scores of the six R&D programs. The final ranking of the six R&D programs is E, D > C > B > A, F in order.

As stated before, the primary reason for using DEA for comparing the performance of R&D programs with different objectives is because DEA is capable of mirroring R&D programs' unique characteristics by assigning relatively high weights to the outputs in which each project has strength. Thus, the magnitude of the weight expresses how highly the variable is evaluated, relatively speaking

Table 4
Results of evaluation and comparison of program performance

Program	Number of projects	Average efficiency score	Mean rank
A	436	0.4434	239.88
B	46	0.6971	366.99
C	21	0.7989	423.95
D	28	0.9257	475.73
E	13	0.9862	475.81
F	4	0.2651	136.75
Relative comparison using Mann–Whitney U test: E, D > C > B > A, F			

$\chi^2 = 124.795$, df = 5, $p = 0.000$.

(Cooper et al., 2000). To examine the relative importance of output variables under each program, the output weight vectors for each project were normalized by dividing each element by their sum. Table 5(a) summarized the average normalized weights of output variables for each program. The numbers in parentheses are the rankings of the weights in each program. Kruskal–Wallis test was also run for differences among the weights of the six programs. It is shown that the weights given to output variables significantly differ across the six programs at the 0.05 level, except u_3 and u_7. The weights seem to reflect each program's own objectives well. In Program A and E whose main objective is to develop human resources, O9(Master's degree students) and O10 (Doctoral degree students) obtained high weights. Program C focuses on basic research so that the variables related with papers, O1, O3, and O4, are relatively highly evaluated. The two programs for applied research, Program D and F, have high weights in the variables on patents, particularly, O5 and O7. As the purpose of Program B is general rather than specific, its weights are not characteristic.

As mentioned in Section 2, however, the weights may not be unique for efficient projects so that it could affect the above interpretations. Thus, the average weights were calculated only for the 476 inefficient projects, as shown in Table 5(b). Some differences in the relative importance of the output variables are found between Table 5(a) and (b). However, in the case of the variables related with the above interpretations, there are not big differences in the rankings of corresponding weights: no or one level change, except for $u_4(3 \rightarrow 6)$ in Program C. On the other hand, for some variables, the weights in Table 5(b) better capture the characteristics of programs (e.g. $u_6(6 \rightarrow 3)$ in Program D). In either case, it is shown that DEA is capable of capturing R&D programs' unique characteristics or strengths well due to its weight flexibility.

4.2. Incorporating relative importance of variables

Although DEA was successfully employed for performance comparison among R&D program, the above analysis has a limitation: unrealistic weights. The weight flexibility is valid for different types of variables, but not for the same types of variables since universal agreements exist on the relative importance of the same types of variables for any program. For example, it is obvious that a paper published in an international journal listed on SCI is more valuable than a paper published in a domestic journal not listed on SCI.

Two alternatives can be introduced to solve the problem. The first one is to aggregate the same types of variables into a single variable. The ten output variables can be categorized into three types: papers, patents, and human resources. The relative importance of the same types of variables can be captured by integrating them with fixed weights. The pairwise comparison method in the analytic hierarchy process (AHP) was employed to derive the priority weights of variables. To avoid confusion between priority weights in the AHP and weights as multipliers in DEA, from now on, the former is called 'priorities', and the latter is described as 'weights'. The relative importance values are determined on a scale of 1–9, where a score of 1 indicates equal importance between the two elements and 9 represents the extreme importance of one element compared with the other one (Saaty, 1980). A reciprocal value is assigned to the inverse comparison; that is, $a_{ji} = 1/a_{ij}$ where a_{ij} denotes the importance of the ith element compared with the jth element. Also, $a_{ii} = 1$ is preserved in the pairwise comparison matrix. Then, the eigenvector method is employed to obtain the priority vectors for each pairwise comparison matrix.

The pairwise comparisons were conducted by eight evaluators who were R&D program managers or the persons concerned in the foundation. The geometric mean was employed for group

Table 5
Average normalized weights of output variables for each program

Program	u_1	u_2	u_3	u_4	u_5
(a) All projects					
A	0.0756 (5)	0.1197 (4)	0.1433 (3)	0.0431 (8)	0.0606 (6)
B	0.1932 (2)	0.0278 (9)	0.1007 (5)	0.0761 (7)	0.0605 (8)
C	0.2771 (1)	0.0192 (9)	0.1929 (2)	0.1831 (3)	0.0599 (6)
D	0.2094 (2)	0.1067 (5)	0.0374 (9)	0.0086 (10)	0.2205 (1)
E	0.0430 (8)	0.0672 (7)	0.0741 (5)	0.0212 (10)	0.0240 (9)
F	0.1258 (4)	0.0000 (7)	0.0709 (6)	0.1465 (3)	0.3753 (1)
Sig.	0.000	0.000	0.290	0.000	0.000
	u_6	u_7	u_8	u_9	u_{10}
A	0.0210 (9)	0.0539 (7)	0.0200 (10)	0.2830 (1)	0.1799 (2)
B	0.1025 (4)	0.0121 (10)	0.1227 (3)	0.0876 (6)	0.2167 (1)
C	0.0117 (10)	0.0575 (7)	0.0688 (5)	0.0555 (8)	0.0741 (4)
D	0.0866 (6)	0.1150 (3)	0.0662 (7)	0.0417 (8)	0.1080 (4)
E	0.0746 (4)	0.0889 (3)	0.0738 (6)	0.0956 (2)	0.4375 (1)
F	0.0715 (5)	0.2101 (2)	0.0000 (7)	0.0000 (7)	0.0000 (7)
Sig.	0.004	0.348	0.000	0.000	0.019
(b) Inefficient projects only					
	u_1	u_2	u_3	u_4	u_5
A	0.0680 (5)	0.1143 (4)	0.1506 (3)	0.0439 (8)	0.0583 (6)
B	0.1383 (3)↓1	0.0093 (10)↓1	0.1116 (5)	0.0640 (7)	0.0519 (8)
C	0.2688 (2)↓1	0.0051 (10)↓1	0.3171 (1)↑1	0.0754 (6)↓3	0.0774 (5)↓1
D	0.2319 (2)	0.0416 (7)↓2	0.0287 (9)	0.0000 (10)	0.2792 (1)
E	0.0559 (7)↑1	0.0806 (5)↑2	0.0963 (4)↑1	0.0275 (9)↑1	0.0312 (8)↑1
F	0.1258 (4)	0.0000 (7)	0.0709 (6)	0.1465 (3)	0.3753 (1)
	u_6	u_7	u_8	u_9	u_{10}
A	0.0191 (9)	0.0534 (7)	0.0173 (10)	0.2938 (1)	0.1813 (2)
B	0.0830 (6)↓2	0.0159 (9)↑1	0.1548 (2)↑1	0.1151 (4)↑2	0.2563 (1)
C	0.0223 (8)↓ 2	0.0188 (9)↓2	0.0840 (4)↑1	0.1060 (3)↑5	0.0251 (7)↓3
D	0.1422 (3)↑3	0.1019 (4)↓1	0.0333 (8)↓1	0.0460 (6)↑2	0.0954 (5)↓1
E	0.0765 (6)↓2	0.1156 (2)↑1	0.0000 (10)↓4	0.1092 (3)↓1	0.4071 (1)
F	0.0715 (5)	0.2101 (2)	0.0000 (7)	0.0000 (7)	0.0000 (7)

decision making. The derived matrices are attached in Appendix B, and the obtained priorities are shown in Table 6. The value of an integrated variable was calculated as the weighted sum of the observed values of each variable. Then, DEA can be conducted with the two original input variables and the three integrated output variables. Regardless of the integration of output variables, the weight flexibility works among the three integrated variables; therefore, the advantage of DEA for R&D program evaluation still holds true.

The other alternative is to apply the AR model for weights restriction. The relative importance can be reflected by placing restrictions for the relationships between the weights of variables. The upper and lower bound for AR1 restrictions can also be derived from the pairwise comparisons. Contrary to the first alternative where the geometric mean is used to aggregate multiple evaluators' judgment, the bounds are determined by using individual evaluator's priorities as

$$L_{ij} = \min_k \frac{w_{ki}}{w_{kj}}, \quad U_{ij} = \max_k \frac{w_{ki}}{w_{kj}}, \tag{5}$$

where w_{ki} indicates the priority score of u_i assigned by kth evaluator. This method for deriving the bounds has often been used in the previous studies (Shang and Sueyoshi, 1995; Takamura and Tone, 2003). The resulting restrictions are as follows:

$$\begin{aligned}
&2.2 \leqslant \frac{u_1(\text{Domestic SCI papers})}{u_2(\text{Domestic non-SCI papers})} \leqslant 5,\\
&1.8 \leqslant \frac{u_3(\text{International SCI papers})}{u_4(\text{International non-SCI papers})} \leqslant 4,\\
&1 \leqslant \frac{u_3(\text{International SCI papers})}{u_1(\text{Domestic SCI papers})} \leqslant 2,\\
&1.5 \leqslant \frac{u_6(\text{Domestic granted patents})}{u_5(\text{Domestic applied patents})} \leqslant 3.5,\\
&1 \leqslant \frac{u_8(\text{Foreign granted patents})}{u_7(\text{Foreign applied patents})} \leqslant 2.8,\\
&1 \leqslant \frac{u_8(\text{Foreign granted patents})}{u_6(\text{Domestic granted patents})} \leqslant 4,\\
&1 \leqslant \frac{u_{10}(\text{Doctoral degree students})}{u_9(\text{Master's degree students})} \leqslant 5.
\end{aligned} \tag{6}$$

These constraints are incorporated into the basic model. It is shown that restrictions are made only for the relationships among the same type of variables. Thus, the weight flexibility still exists among the different types of variables. What should be noted here is, in the presence of the weights restriction, infeasible solutions occur for the DMUs that have the observed value of 0 in the variables included in the constraints. To solve this problem, the zero values were replaced by 0.00000001.

To illustrate the effects of the two alternatives, Table 7 compares the efficiency scores of 28 projects in Program D for the two cases with the basic BCC efficiency. There exist notable differences in the efficiency scores among the three cases. Many of the efficient DMUs in the basic model were found to be inefficient in the other two cases. What is prominent is sharp falls in the rankings of D5, D16, and D21. On the other hand, some of the DMUs are more highly ranked in the two alternatives (e.g. D8, D14, D22, D23). It is concluded that two alternatives can produce more realistic results by mirroring the relative importance of output variables. The portion of the efficient DMUs is also decreased from 75.00% (basic) to 42.86% (AR) and 35.71% (output integration), which indicates the two alternatives make a sharper discrimination among DMUs than the basic BCC model.

The program-level comparisons were made in the same way. The two alternatives were applied to the basic model, the output-oriented BCC model for comparison of efficiency between different systems model used in 4.1. Kruskal–Wallis test with a post hoc Mann–Whitney U test was then also run to compare performance of the six R&D programs. Table 8 shows the results with the previous ones obtained from the basic model. The difference between the basic evaluation and the two alternatives is not so significant except Program E. Program E is ranked at top in the basic model while it is fourth and fifth in the AR model and output integration, respectively. This is because the projects of Program E has relatively high values for the variables that are considered less important in the pairwise comparisons such as O2 (Domestic non-SCI papers) and O5 (Domestic applied patents). Program E benefited from the full flexibility of the weights selection in the basic model, but it turned out to be a bad performer under more realistic weights. A marked difference between the AR model and output integration is the performance of Program F. It shows better performance in the output integration than in the AR model in terms of mean rank. The plausible explanation for this is the fixed priorities of the output variables led to sharp falls of the efficiency scores of the projects belonging to Program A and E, which in turn made increases in mean rank of Program F whose number of projects is very small. Although it is hard to make judgment about which of the two alternatives is better due to the existence of their pros and cons, what is evident is the performance comparison in which the relative

Table 6
Priorities of output variables for integration

Integrated outputs	Output variables	Priorities
Papers	Domestic SCI papers	0.27
	Domestic non-SCI papers	0.07
	International SCI papers	0.52
	International non-SCI papers	0.14
Patents	Domestic applied patents	0.06
	Domestic granted patents	0.26
	Foreign applied patents	0.12
	Foreign granted patents	0.56
Human resources	Master's degree students	0.25
	Doctoral degree students	0.75

Table 7
Efficiency scores of projects in Program D for three cases

Projects	Basic		AR		Output integration	
	Score	Rank	Score	Rank	Score	Rank
D1	1.0000	1	0.9875	13	0.8258	18
D2	1.0000	1	1.0000	1	1.0000	1
D3	1.0000	1	1.0000	1	1.0000	1
D4	1.0000	1	1.0000	1	1.0000	1
D5	1.0000	1	0.8915	20	0.7884	20
D6	1.0000	1	1.0000	1	1.0000	1
D7	1.0000	1	1.0000	1	1.0000	1
D8	0.4756	28	0.4437	26	0.4309	26
D9	1.0000	1	1.0000	1	1.0000	1
D10	1.0000	1	0.8987	17	0.9341	13
D11	1.0000	1	0.8923	19	0.8527	17
D12	1.0000	1	0.9547	14	0.8632	16
D13	1.0000	1	0.8978	18	0.9151	14
D14	0.6084	27	0.5977	24	0.5964	22
D15	1.0000	1	1.0000	1	0.9999	11
D16	1.0000	1	0.6400	22	0.5558	24
D17	0.7460	24	0.4196	28	0.4189	27
D18	0.7670	23	0.6002	23	0.5895	23
D19	0.7292	25	0.4261	27	0.4002	28
D20	1.0000	1	0.9076	15	0.6965	21
D21	1.0000	1	0.8204	21	0.8031	19
D22	0.9790	22	0.9019	16	0.8769	15
D23	0.6450	26	0.5463	25	0.4773	25
D24	1.0000	1	1.0000	1	1.0000	1
D25	1.0000	1	1.0000	1	1.0000	1
D26	1.0000	1	1.0000	1	1.0000	1
D27	1.0000	1	1.0000	1	0.9998	12
D28	1.0000	1	1.0000	1	1.0000	1
Number of efficient units	21 (75.00%)		12 (42.86%)		10 (35.71%)	

Table 8
Comparison of program performance among the three cases

Program	Number of projects	Basic		AR		Output integration	
		Average efficiency score	Mean rank	Average efficiency score	Mean rank	Average efficiency score	Mean rank
A	436	0.4434	239.88	0.3055	244.42	0.0773	229.77
B	46	0.6971	366.99	0.5793	406.07	0.2157	421.28
C	21	0.7989	423.95	0.5762	406.71	0.5648	493.67
D	28	0.9257	475.73	0.7007	460.88	0.6411	504.64
E	13	0.9862	475.81	0.2875	262.54	0.0556	370.00
F	4	0.2651	136.75	0.1100	80.00	0.0749	389.75
Relative comparison		E, D > C > B > A, F		D > C, B > E, A > F		D, C > B > F > E > A	

χ^2 = 124.8, df = 5, p = 0.000; χ^2 = 107.0, df = 5, p = 0.000; χ^2 = 108.6, df = 5, p = 0.000.

importance of the same types of variables is mirrored is more realistic and reasonable than when this is not the case.

5. Conclusions

We measured and compared the performance of the six national R&D programs with heterogeneous objectives using DEA. Every project in every program was evaluated together, and Kruskal–Wallis test with a post hoc Mann–Whitney *U* test was then run to compare performance of R&D programs. The two alternative approaches to incorporating the importance of variables in reality, the AR model and output integration, were also considered.

Due to the heterogeneity of national R&D programs' objectives, few studies have been conducted on the performance comparison among programs. This study contributes to the field in that it filled the void by applying DEA to the national R&D programs. DEA, particularly the model for comparison of efficiency between different systems, was proved to be effective for performance comparison among R&D programs with heterogeneous objectives. The "benefit of the doubt" weighting procedure of DEA enabled R&D programs' unique characteristics or strengths to be captured. Nobody can complain about the results since each program is evaluated under its most favorable setting.

The DEA results are expected to provide practical implications for policy making on national R&D programs. The limited resources can be effectively allocated to several R&D programs based on their performance rankings. R&D programs doing well (e.g. Program C and D) deserve more investments; on the other hand, poor programs (e.g. Program A and F) have to be terminated or funds given to them should be cut down unless their performance is improved. Basically, DEA offers the way of improving efficiency for inefficient DMUs although it is not explicitly dealt with in this study. Each inefficient project is provided with the reference set consisting of efficient projects for benchmarking, which in turn results in performance improvement of programs. However, what DEA tells us is the way of improving efficiency, that is, how many outputs should be increased to achieve 100% efficiency, not the way of increasing actual outputs at the current setting. To seek the way of enhancing performance, the reasons for poor performance should be uncovered by examining the context in which poor programs are formulated and implemented, such as project selection procedure, operational regulation, funding systems, etc. It is obvious that the prerequisite for this is to be able to measure and compare the performance of various R&D programs, which is the primary contribution of this study.

Nevertheless, this study is subject to some limitations. Firstly, since the projects that have not been finished at the time of data collection were not included, program performance was measured without them. Secondly, despite the fact that it takes several years for R&D outputs to be achieved, the outputs produced only for two years after termination of projects were considered. These limitations will be overcome if the analysis is conducted again at some time in future. Thirdly, it may occur that a R&D program is considered as a high-performer, even though they failed to achieve its own objectives, but accomplished excellent outcomes in another area. Although it was not found in this study, when this is the case, judgment could be controversial. These issues should be dealt with in future research. Another fruitful avenue for future research is to employ various types of extended DEA models and compare the results. Considering another model is expected to lead us to seek a better way of evaluating and comparing the performance of national R&D programs with heterogeneous objectives.

Appendix A. Descriptive statistics of data

Program	Statistics	I1[a]	I2	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
Program A	Mean	256.79	2.63	0.37	2.79	3.34	0.33	0.42	0.21	0.06	0.06	3.03	0.56
	Stdv	60.09	0.71	1.07	3.99	4.40	1.08	1.14	1.17	0.36	0.70	3.13	0.91
	Max	360	5	10	27	45	13	11	17	4	10	19	5
	Min	113	1	0	0	0	0	0	0	0	0	0	0
Program B	Mean	4891.59	1.00	3.48	2.70	60.61	0.30	5.15	4.11	5.33	1.87	10.87	5.24
	Stdv	1315.41	0.00	5.21	7.13	44.60	0.89	7.04	5.00	9.40	3.12	9.01	4.45
	Max	7905	1	20	33	226	5	27	17	45	18	38	20
	Min	2418	1	0	0	7	0	0	0	0	0	0	0
Program C	Mean	5763.67	11.57	18.48	16.10	216.81	10.24	16.10	6.00	5.76	1.67	97.05	29.10
	Stdv	1626.78	6.56	16.91	15.32	140.19	15.27	16.03	7.70	14.62	3.06	63.30	13.65
	Max	8164	28	67	56	630	55	57	28	68	12	330	54
	Min	690	5	0	0	58	0	0	0	0	0	30	7
Program D	Mean	6528.68	14.30	17.43	145.61	258.32	22.68	43.93	24.21	11.36	6.11	170.96	42.00
	Stdv	1245.78	4.82	17.01	146.45	192.25	33.11	32.18	16.07	14.12	6.31	102.08	28.74
	Max	9233	22	67	574	897	140	135	71	69	21	376	106
	Min	4638	6	0	8	47	0	0	0	0	0	55	9

(*continued on next page*)

Appendix A (*continued*)

Program	Statistics	I1[a]	I2	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
Program E	Mean	1300.00	10.92	0.85	9.77	32.08	0.85	2.38	1.31	0.62	0.31	9.15	5.46
	Stdv	212.22	3.86	1.28	6.31	15.47	1.21	2.18	1.70	1.12	0.85	5.49	4.29
	Max	1600	18	4	21	74	3	8	6	3	3	20	16
	Min	1035	6	0	1	10	0	0	0	0	0	3	0
Program F	Mean	4500.00	29.00	3.75	3.00	46.75	1.75	5.50	3.25	1.50	0.00	15.00	4.25
	Stdv	1154.70	2.71	4.99	2.16	24.34	2.87	4.04	3.95	1.91	0.00	4.76	3.30
	Max	5500	31	11	5	71	6	9	9	4	0	18	8
	Min	3500	25	0	0	13	0	0	0	0	0	8	0

[a] Unit: one million won.

Appendix B. Pairwise comparison matrices for output variables

B.1. Pairwise comparison matrix for paper-related variables

	O1: Domestic SCI papers	O2: Domestic non-SCI papers	O3: International SCI papers	O4: International non-SCI papers	Priorities
O1: Domestic SCI papers	1	4	1/2	2	0.27
O2: Domestic non-SCI papers	4	1	1/7	1/2	0.07
O3: International SCI papers	2	7	1	4	0.52
O4: International non-SCI papers	1/2	2	1/4	1	0.14

B.2. Pairwise comparison matrix for patent-related variables

	O5: Domestic applied patents	O6: Domestic granted patents	O7: Foreign applied patents	O8: Foreign granted patents	Priorities
O5: Domestic applied patents	1	1/5	1/3	1/7	0.06
O6: Domestic granted patents	5	1	3	1/3	0.26
O7: Foreign applied patents	3	1/3	1	1/5	0.12
O8: Foreign granted patents	7	3	5	1	0.56

B.3. Pairwise comparison matrix for human resources-related variables

	O9: Master's degree students	O10: Doctoral degree students	Priorities
O9: Master's degree students	1	1/3	0.25
O10: Doctoral degree students	3	1	0.75

References

Adams, J.D., Grilliches, Z., 2000. Research productivity in a system of universities. In: Encaoua, D. (Ed.), The Economics and Econometrics of Innovation. Kluwer Academic Publisher, Boston.

Allen, R., Athanassopoulos, A., Dyson, R.G., Thanassoulis, E., 1997. Weights restrictions and value judgments in data envelopment analysis: Evolution, development and future directions. Annals of Operations Research 73, 13–34.

Banker, R.D., Charnes, A., Cooper, W.W., 1984. Some models for estimating technical and scale inefficiency in data envelopment analysis. Management Science 30 (9), 1078–1092.

Banker, R.D., Charnes, A., Cooper, W.W., Swarts, J., Thomas, D.A., 1989. An introduction to data envelopment analysis with some of its models and their uses. Research in Governmental and Nonprofit Accounting 5, 125–163.

Bound, J., Cumins, C., Griliches, Z., Hall, H.H., Jaffe, A., 1984. R&D, Patent, and Productivity. University of Chicago Press, Chicago.

Boussofiane, A., Dyson, R.G., Thanassoulis, E., 1991. Applied data envelopment analysis. European Journal of Operational Research 52 (1), 1–15.

Brinn, A., Jones, M.J., Pendlebury, M., 1996. UK accountants' perceptions of research journal quality. Accounting and Business Research 26 (3), 265–278.

Charnes, A., Cooper, W.W., Rhodes, E., 1978. Measuring efficiency of decision making units. European Journal of Operational Research 2 (6), 429–444.

Cherchye, L., Moesen, W., Rogger, N., van Puyenbroeck, T., 2007. An introduction to 'benefit of the doubt' composite indicators. Social Indicators Research 82 (1), 111–145.

Cooper, W.W., Seiford, L.M., Tone, K., 2000. Data Envelopment Analysis: Theory, Methodology, and Applications, References and DEA-Solver Software. Kluwer Academic, Boston.

Dyson, R.G., Thanassoulis, E., 1988. Reducing weight flexibility in date envelopment system. The Journal of Operational Research Society 39 (6), 563–576.

Eilat, H., Golany, B., Shtub, A., 2006. Constructing and evaluating balanced portfolios of R&D projects with interactions: A DEA based methodology. European Journal of Operational Research 172 (3), 1018–1039.

Feng, Y.J., Lu, H., Bi, K., 2004. An AHP/DEA method for measurement of the efficiency of R&D management activities in universities. International Transactions in Operational Research 11, 181–191.

Garg, K.C., Gupta, B.M., Jamal, T., Roy, S., Kumar, S., 2005. Assessment of impact of AICTE funding on R&D and educational development. Scientometrics 65 (2), 151–160.

Graves, S.B., Langowitz, N.S., 1996. R&D productivity: A global multi-industry comparison. Technology Forecasting and Social Change 53 (2), 125–137.

Kocher, M.G., Luptacik, M., Sutter, M., 2006. Measuring productivity of research in economics: A cross-country study using DEA. Socio-Economic Planning Sciences 40 (4), 314–332.

Korhonen, P., Tainio, R., Wallenius, J., 2001. Value efficiency analysis of academic research. European Journal of Operational Research 130 (1), 121–132.

Lee, H., Park, Y., 2005. An international comparison of R&D efficiency: DEA approach. Asian Journal of Technology Innovation 13 (2), 207–221.

Lee, M., Son, B., Om, K., 1996. Evaluation of national R&D projects in Korea. Research Policy 25 (5), 805–818.
Nederhof, A.J., van Raan, A.F.J., 1987. Peer review and bibliometric indicators of scientific performance: A comparison of cum laude doctorates with ordinary doctorates in physics. Scientometrics 11, 330–350.
Nederhof, A.J., van Raan, A.F.J., 1993. A bibliometric analysis of six economics research groups: A comparison with peer review. Research Policy 22 (4), 353–368.
OECD, 2001. OECD Science, Technology, and Industry Scoreboard. OECD, Paris.
Paradi, J.C., Smith, S., Schaffnit-Chatterjee, C., 2002. Knowledge worker performance analysis using DEA: An application to engineering design teams at Bell Canada. IEEE Transactions on Engineering Management 49 (2), 161–172.
Rosen, D., Schaffnit, C., Paradi, J.C., 1998. Marginal rates and two-dimensional level curves in DEA. Journal of Productivity Analysis 9, 205–232.
Saaty, T., 1980. The Analytic Hierarchy Process. McGraw-Hill, New York.
Scherer, F.M., 1983. The propensity to patent. International Journal of Industrial Organization 1 (1), 107–128.
Shang, J., Sueyoshi, T., 1995. A unified framework for the selection of a flexible manufacturing system. European Journal of Operational Research 85 (2), 297–315.
Swink, M., Talluri, S., Pandejpong, T., 2006. Faster, better, cheaper: A study of NPD project efficiency and performance tradeoffs. Journal of Operations Management 24 (5), 542–562.
Takamura, Y., Tone, K., 2003. A comparative site evaluation study for relocating Japanese government agencies out of Tokyo. Socio-Economic Planning Sciences 37 (2), 85–102.
Thompson, R.G., Langemeir, L.N., Lee, C., Lee, E., Thrall, R.M., 1990. The role of multiplier bounds in efficiency analysis with application to Kanssa farming. Journal of Econometrics 1 (2), 93–108.
Wang, E.C., Huang, W., 2007. Relative efficiency of R&D activities: A cross-country study accounting for environmental factors in the DEA approach. Research Policy 36 (2), 260–273.
Zhang, A., Zhang, Y., Zhao, R., 2003. A study of the R&D efficiency and productivity of Chinese firms. Journal of Comparative Economics 31 (3), 444–464.