Chapter 12
Applications of Data Envelopment Analysis in Education

Emmanuel Thanassoulis, Kristof De Witte, Jill Johnes, Geraint Johnes, Giannis Karagiannis, and Conceição S. Portela

Abstract Non-parametric methods for efficiency evaluation were designed to analyse industries comprising multi-input multi-output producers and lacking data on market prices. Education is a typical example. In this chapter, we review applications of DEA in secondary and tertiary education, focusing on the opportunities that this offers for benchmarking at institutional level. At secondary level, we investigate also the disaggregation of efficiency measures into pupil-level and school-level effects. For higher education, while many analyses concern overall

Part of this research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) – Research Funding Program: THALES. Investing in Knowledge Society through the European Social Fund. The views expressed in this chapter are those of the authors and no representation is made that they are shared by any funding body.

E. Thanassoulis (✉)
Aston Business School, Aston University, Birmingham, UK
e-mail: e.thanassoulis@aston.ac.uk

K. De Witte
Leuven Economics of Education Research, Faculty of Economics and Business,
KU Leuven, Leuven, Belgium

Top Institute for Evidence Based Education Research, Maastricht University,
Maastricht, The Netherlands

J. Johnes
Huddersfield University Business School, University of Huddersfield, Huddersfield, UK

G. Johnes
Lancaster University Management School, Lancaster University, Lancaster, UK

G. Karagiannis
Department of Economics, University of Macedonia, Thessaloniki, Greece

C.S. Portela
Católica Porto Business School, Porto, Portugal

J. Zhu (ed.), *Data Envelopment Analysis*, International Series in Operations
Research & Management Science 238, DOI 10.1007/978-1-4899-7684-0_12

institutional efficiency, we examine also studies that take a more disaggregated approach, centred either around the performance of specific functional areas or that of individual employees.

Keywords DEA • Efficiency • Education • Benchmarking • Pupil-level effects • School efficiency • Higher education efficiency

12.1 Introduction

Data envelopment analysis (DEA) was originally developed to provide a means of efficiency evaluation in the context of 'not-for-profit entities participating in public programs' (Charnes et al. 1978). Not all such entities are providers of public goods – in the sense that their output is non-rival and non-excludable – but they are all characterised by a production process that converts a multiplicity of inputs into a multiplicity of outputs for which market prices are absent. In this respect, education represents a classic example of a sector which is well served by DEA.

Beyond elementary education, schools, colleges and universities provide specialist tuition in a wide variety of subject areas. Inputs include the expertise of teachers in each of these subjects, and the extent of their specialism means that teaching in each subject (or at least in each cluster of subjects) should be regarded as a distinct input. Likewise, a multiplicity of outputs reflects students who (at secondary level) might take a vocational or academic route, or who (at tertiary level) specialise in particular disciplines. For many higher education institutions, research provides a further distinct output. There exists considerable synergy between the various activities, and cross-subsidisation is common.

The education sector in most countries comprises both public and private provision. In the private sector as much as in the public, joint production of multiple outputs is common. Given unobserved heterogeneity across students, assessing the benefit of education for any individual student is problematic, and in consequence the price charged for educational services does not have many of the characteristics that are associated with market pricing. The presence of multiple inputs, multiple outputs, and prices that are unlikely to serve a useful purpose as weights, are features that combine to make DEA an instructive tool in this context.

The structure of this chapter is as follows. In the next section, we look at applications of DEA in secondary education. These include studies that operate at various levels of aggregation. First we review studies that focus on relatively highly aggregated data, evaluating the efficiency of schools. Typically these include as inputs the characteristics of schools and their pupil intakes; meanwhile outputs include various measures of pupil attainment. A particularly interesting development in this area is the emergence of online platforms that allow schools to enter data about their own performance and then compare this performance with that of peers (who have likewise entered data onto the platform). This illustrates very vividly the scope for DEA and similar methodologies to provide benchmarking information that can be useful in the dissemination of good practice and hence in the process of securing efficiency improvements. We next proceed to review studies that focus on the pupil as the level of analysis. These are relatively rare, partly because of data availability issues, but partly also because the computational burden of DEA becomes considerable when dealing with large numbers of individual pupils. However these studies are insightful, not least because they allow each school's frontier to be understood as an envelope around its students' performance. Our assessment of the relative performance of two otherwise identical students attending different schools might be conditioned by information about differences in the frontiers associated with the schools that they attend. This separation into a school effect and a pupil effect has some aspects in common with the statistical approach of multilevel modelling, and we draw comparisons.

In reviewing the literature on school efficiency, we consider also the way in which efficiency has been observed to change over time, and in particular focus on the extent to which such change is due to either change in the distribution of efficiencies or movements of the efficiency frontier itself.

In Sect. 12.3, we review studies of DEA as applied to higher education institutions, focusing on analyses that use data aggregated to the level of the institution. These studies include analyses of the cost efficiency and technical efficiency of overall operations. We also consider studies that have focused on particular aspects of university activities, specifically including investigation of the efficiency of administrative services and the efficiency of research production in the university sector.

The basic DEA model has been extended to study a variety of applications in higher education, and some of these are reviewed in this section of the chapter. For instance, there are some examples of merger activity in higher education – (concern about) efficiency is often cited as a cause and (a change in) efficiency is an effect of this activity. Higher education is also one of the areas in which network DEA has been applied, yielding results that are informative about the sources of relative inefficiency within the 'black box' of production.

More disaggregated data have been used in the higher education sector to evaluate the performance of individual staff members along a number of dimensions. We review these studies in Sect. 12.4, focusing specifically first on academics' research output, and second on their teaching output.

The chapter ends with a conclusion that draws together the main threads of the preceding discussion.

12.2 Applications of DEA in Secondary Education

12.2.1 Introduction

Educational data follow typically a hierarchical structure, since pupils are nested within classes, classes are nested within schools, schools are nested within districts, school districts are nested within states, which are nested within countries. Within each of these levels there are variables of interest for educational policies (e.g. pupils' socio-economic background, ability of peers in the same class, size of the school, state policies regarding the autonomy of schools etc.). The analysis of data at several levels requires the adaptation of existing parametric or non-parametric techniques which are usually designed for single level analysis.

Since pupil-level data are at the lowest disaggregated level, the analysis of these data can provide more information to practitioners and researchers. As a result, most school effectiveness research is actually undertaken at this level of analysis. Research on school effectiveness started in the 60s. Its most influential article is the controversial Coleman Report (1966). Conclusions of this report pointed to the lack of importance of the schools themselves in explaining attainment by pupils. This, being a counter intuitive finding, gave rise to a number of studies whose aim was to prove that schools did make a difference. These studies typically approach education as a production process, where student outcomes are a function of several variables. The 'educational production function' framework groups inputs into four main dimensions: Family background, Peer influences; School inputs; and Innate abilities of students (see Hanushek 1979). The variables considered within this general model are constructs that need to be operationalized using specific quantitative measures.[1] For example, for operationalizing the innate abilities prior attainment before entering a given stage of education under assessment is usually used as a proxy. The use of prior attainment to explain subsequent attainment gave rise to what have been termed value-added (**VA**) studies. Such studies differ from other school effectiveness studies as they focus on the progress schools help pupils to make relative to their different starting points (see e.g. Meyer 1997; Goldstein et al. 2000). As a result, VA studies have a longitudinal perspective of pupil attainment: they are usually undertaken at the pupil-level and consider the achievements of pupils over a certain cycle of studies, with achievements on entry and on exit of that cycle as the main variables of interest. For detailed discussions around VA models and implementations see OECD (2008). Clearly several other variables can be considered within the analysis. Typically socio-economic characteristics of the students are considered as an additional and very important driver of exit achievement. Note however, that consideration of prior (entry) achievement "implicitly controls for socio-economic status and other background factors to the extent that their influence on the post-test is already reflected in the pre-test score" (Ballou et al. 2004, p. 38).

Assessments of the comparative efficiency of secondary schools in value-added fall into two broad categories in terms of the data used. Those using aggregate and those using pupil-level data. The type of data used is a combination of data availability and the issues to be addressed through the assessment. We outline sample applications from both genres. However, in the context of DEA, earlier

[1] Clearly qualitative studies are of utmost importance in education, but we address here only quantitative studies.

applications used aggregate pupil data (i.e. school data or school districts data) and so we look first at DEA applications using these types of data. Note that within this type of studies a schools' VA perspective can be taken, but an efficiency perspective may also be investigated. Banker et al. (2004) distinguished between the assessment of efficiency and effectiveness in schools, which could be operationalized by choosing different types of variables for the assessment. Mayston (2003) considers a similar classification of school studies, where the distinction lies in the consideration or not of expenditures and school resources. In this chapter we take efficiency to be 'value for money' (i.e. where school resources including expenditures are central), while effectiveness is value-added (i.e. where school resources and other 'endogenous' variables are not permitted to explain exit attainment). In this respect we adopt the Mayston (2003) distinction between efficiency and effectiveness in education assessments.

12.2.2 Applications Using Aggregate Pupil Data

Aggregate pupil data are more readily amenable to DEA than pupil-level data and this explains in large measure why DEA assessments initially relied on such data. Pupil-level data, as we will see later, offer greater scope for addressing questions such as identifying pupil-level as distinct from school-level effects on VA. Nevertheless, the readily available aggregate data also make it possible to address significant questions about school effectiveness and efficiency as we now illustrate.

12.2.2.1 An Overview of DEA Applied to Aggregate Pupil Data

The number of DEA applications on aggregate pupil data has grown since the 80s when the first studies of this type emerged. The first school efficiency study was that of Bessent and Bessent (1980), followed by a very influential paper of Charnes et al. (1981) that will be covered in some detail below. An extensive survey of the earlier literature is provided by De Witte and López-Torres (2015).

The DEA studies that have used the school as the level of analysis have used in general standard DEA models (except in a few cases), differing mainly on the type of inputs and outputs used, and therefore on the focus of the analysis. A general consensus regarding the type of inputs that should be considered in such studies has emerged in the literature as three groups of variables are usually considered: (i) those reflecting characteristics of pupils (like prior attainment and socio-economic characteristics), (ii) those reflecting characteristics of the school (like number of teaching and non-teaching staff, expenditure per pupil, size of school, or class size), and (iii) those reflecting characteristics of teachers (like their salary, experience, or level of education). Regarding outputs, the consensus is that these should relate to standardized test scores, but they have been aggregated in several forms like the median (Bessent and Bessent 1980), the mean (Mizala et al. 2002;

Table 12.1 Input/output set in Charnes et al. (1981)

Inputs	Outputs
Maternal education level	Attainment in reading
Highest occupation of a family member	Attainment in maths
Parental visits to school	Self esteem
Number of teachers	

Muñiz 2002), or the proportion of pupils achieving more than a certain grade (Bradley et al. 2001). Other relevant outputs also related to pupils' achievement are the number of approvals or success rates (Kirjavainen and Loikkanen 1998; Muñiz 2002; Oliveira and Santos 2005), attendance rate (Arnold et al. 1996; Bradley et al. 2001), number of graduates (Kirjavainen and Loikkanen 1998), and percentage of students who do not drop out from school (Arnold et al. 1996).

One of the studies that can be classified within a school effectiveness perspective that gave rise to subsequent applications of DEA in education (both at the school and at the pupil-level) is that of Charnes et al. (1981). This application uses DEA in the context of what could be called a 'matched-school' experiment and it has led to what became known as disentangling 'managerial' from 'program' efficiency. The approach spawned a plethora of applications both within and outside education.

The study of Charnes et al. (1981) originated in the USA, where federal authorities at that time wished to test the effectiveness of a program of interventions with primary school children known as 'Program Follow Through' (PFT). The aim of PFT was to compensate disadvantaged children by instituting academic and indeed non-academic interventions (e.g. social, nutritional and other counselling), which would be offered to 'treated' cohorts of children. Each 'treated cohort' was matched with an 'untreated' cohort. The intention of the study was to establish whether PFT was achieving its aim of compensating to some degree for the disadvantaged background of the children concerned.

The study used data both from treated and untreated cohorts, normalised as per 100 pupils. The input/output set used in that study is shown in Table 12.1, where suitably constructed measures for the cohort were chosen.

The key aim of the study was to isolate through DEA any ineffectiveness in implementation of the PFT so that the effectiveness of the program itself could be identified. The approach used is illustrated graphically below.

Let two sets of units (cohorts of pupils in the case of the study under consideration) operate under 'program' 1 and 2 respectively. Assume the units operate under constant returns to scale, using two inputs to secure one output. In Fig. 12.1 ACD is the efficient boundary of program 1 and EFG that of program 2. The efficiency of each unit relative to its own policy boundary is its 'managerial' efficiency.

Thus OB/OJ is the managerial efficiency of school J of Program 1 and OK′/OK that of school K operating within Program 2. To discount managerial inefficiencies and isolate 'program' efficiencies each school is projected to the efficient boundary of its policy as depicted in Fig. 12.2.

EFCD is a global or meta frontier enveloping all programs. OM″/OM′ is the 'program 1' efficiency at the input mix of unit M. The overall inefficiency of unit M

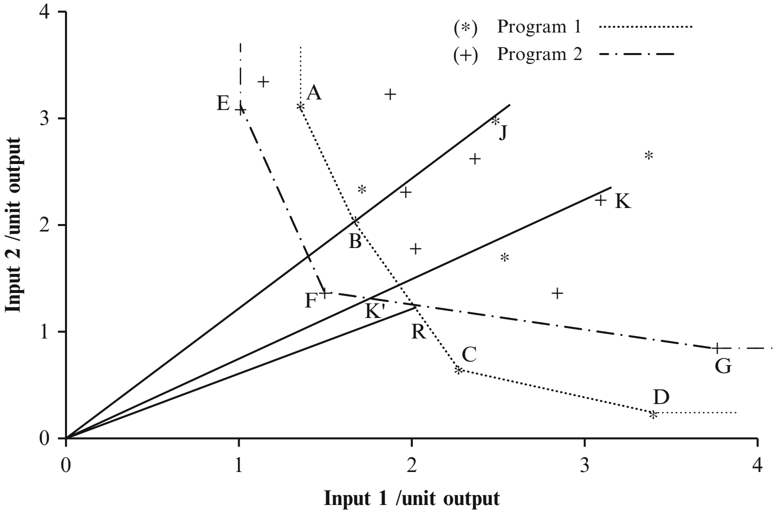

Fig. 12.1 Efficient boundaries drawn by policy

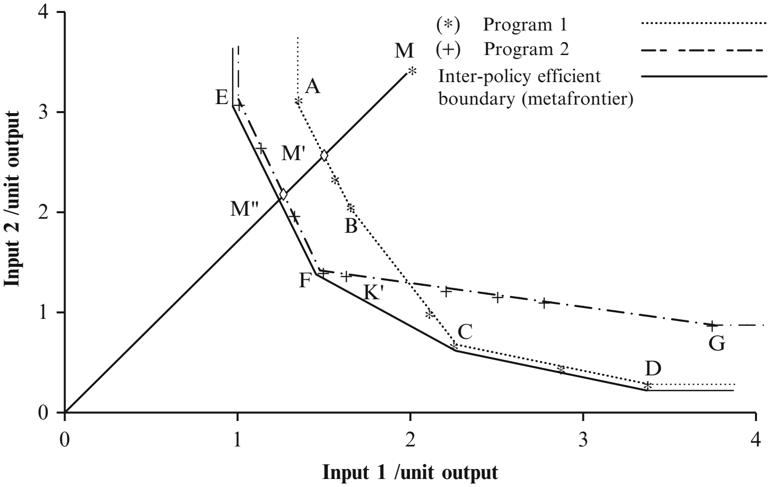

Fig. 12.2 Global (meta) frontier envelops the program boundaries

is OM″/OM. This is decomposed into OM″/OM′ attributable to the program under which unit M operates (Program 1 in this case) and OM′/OM is the component of the inefficiency of Unit M that is attributable to its own management. (Note that though here the overall inefficiency of unit M, OM″/OM is multiplicatively decomposed so that OM″/OM = (OM″/OM′) × (OM′/OM) this will not generally be the case where the efficient projection of a unit such as M is such that there are slacks to be eliminated in order to render that projection Pareto efficient).

The process illustrated above can be operationalised using multiple inputs and outputs by solving the DEA models (12.1) and (12.3) as explained below.

Assume that there are N_p DMUs operating under program p ($p = 1 \ldots P$), and that DMU j of program p uses inputs x_{ij}^{p} (i = 1,. . .,m) to secure outputs y_{rj}^{p} (r = 1,. . .,s). The managerial technical input efficiency of DMU j$_0$ of program p is the optimal value of k_{j0}^{p} in (12.1).

$$\begin{aligned} & Min\, k_{j0}^{p} - \varepsilon\left(\sum\nolimits_i S_i^- + \sum\nolimits_r S_r^+\right) \\ & st: \\ & \sum_{j=1}^{n} \lambda_j x_{ij}^{p} = k_{j0}^{p} x_{ij0}^{p} - S_i^- \quad i = 1, \ldots, m \\ & \sum_{j=1}^{n} \lambda_j y_{rj}^{p} = y_{rj0}^{p} + S_r^+ \qquad r = 1, \ldots, s \\ & \lambda_j \geq 0, \forall j, \quad S_i^-, S_r^+ \geq 0, \quad \forall i, \text{ and } r, k_{j0}^{p} \text{ free} \end{aligned} \tag{12.1}$$

Let ($x_{ij_0}^{tp}$, $i = 1 \ldots m$, $y_{rj_0}^{tp}$, $r = 1 \ldots s$) be a set of input-output levels that would render DMU j_0 Pareto-efficient within its program p. We will use the set ($x_{ij_0}^{tp}$, $i = 1 \ldots m$, $y_{rj_0}^{tp}$, $r = 1 \ldots s$) yielded by model (12.1) so that:

$$\begin{aligned} & x_{ijo}^{tp} = \sum_{j=1}^{n} \lambda_j^* x_{ij}^{p} = k_{jo}^{p*} x_{ijo}^{p} - S_i^{-*} \quad i = 1, \ldots, m \\ & y_{rjo}^{tp} = \sum_{j=1}^{n} \lambda_j^* y_{rj}^{p} = y_{rjo}^{p} + S_r^{+*} \qquad r = 1, \ldots, s \end{aligned} \tag{12.2}$$

where the superscript * denotes the optimal value of the corresponding variable in (12.1).

The program efficiency at the input mix of DMU j$_0$ is the optimal value p$_0$* of p$_0$ in model (12.3).

$$\begin{aligned} & Min \qquad p_0 - \varepsilon\left(\sum\nolimits_i S_i^- + \sum\nolimits_r S_r^+\right) \\ & st: \\ & \sum_{p=1}^{P} \sum_{j=1}^{N_p} \lambda_j x_{ij}^{tp} = p_0 x_{ij0}^{tp} - S_i^- \quad i = 1, \ldots, m \\ & \sum_{p=1}^{P} \sum_{j=1}^{N_p} \lambda_j y_{ij}^{tp} = y_{rj0}^{tp} + S_r^+ \qquad r = 1, \ldots, s \\ & \lambda_j \geq 0, \forall j, \quad S_i^-, S_r^+ \geq 0, \quad \forall i, \text{ and } \quad r, \; p_0 \text{ free} \end{aligned} \tag{12.3}$$

Charnes et al. (1981) used the above approach for illustrative rather than definitive purposes in order to disentangle managerial from program efficiencies in the case of PFT and untreated cohorts labelled NFT (non follow through). They point in particular how ‘inter-program’ areas such as the solid boundary FC in Fig. 12.2 can indicate other amalgams of Programs not constituted initially that can give useful indications of interventions that might be constructed as new Programs.

To the extent that all schools have as basic aim enhancing the educational attainments of pupils but at the same time pupils and/or schools often operate in different contexts (e.g. they may operate under different regimes (e.g. fee charging vs publicly funded schools)) the approach outlined above isolates the impact of context from the effectiveness of the school itself, when we control for context. The approach has been adapted in many educational contexts, e.g. see Portela and Thanassoulis (2001); Thanassoulis and Portela (2002); De Witte et al. (2010); Mancebón et al. (2012). Further, the same approach has been used in many other areas away from education such as in banking (e.g. Golany and Storbeck 1999; Johnes et al. 2014), and water (e.g. De Witte and Marques 2009).

Some studies measuring school effectiveness did not consider the approach of Charnes et al. (1981) but applied in general a non-parametric methodology to the estimation of school effectiveness. An example can be found in Cherchye et al. (2010). These authors used as inputs the total number of instruction units assigned to a particular pupil. For their Flemish application, this consists of regular (REG) and additional, so-called ‘equal educational opportunity’ (EEO), instruction units (depending on certain ‘disadvantageous’ pupil characteristics). Output is defined on the basis of test scores in three dimensions: mathematics, technical reading and writing, collected at the end of the second year.

Regarding school efficiency studies these are distinguished from effectiveness studies by the fact that a central consideration is that of school expenditures on the input side of the assessment. The aim is to assess the extent to which some schools or school districts are more cost effective than others in providing school outcomes. On the input side of such assessments, the inputs relating to prior attainment or to the socio-economic characteristics of pupils may also be considered as they too impact pupil outcomes.

An example of an early study on school efficiency is that by Färe et al. (1989), where the authors considered 40 Missouri school districts, using on the input side variables such as the number of students, the net expenditure and the number of eighth grade teachers while outputs were the number of students passing three types of exams. Another study of the same type is that of Ruggiero (1999) where the author addressed explicitly the cost efficiency of 584 school districts in New York. Conclusions point to the striking figure of 64 % of school districts being cost inefficient, (in Färe et al. (1989) more than half of the school districts were considered efficient). Ruggiero (1999) used a single input in a DEA model (the expenditure per pupil) and considered as environmental variables the percentage of minority students and the percentage of limited English students (a measure of the effect of the environment on costs was estimated in this study). Fukuyama and Weber (2002) also analysed cost efficiency (and allocative efficiency) of 310 Texas school districts using several model specifications. They divided inputs into variable (number of administrators, teachers, teacher aides and support staff) and fixed (predicted achievement of pupils based on their prior achievement and socio-economic characteristics and operating expenses per pupil). Outputs used were the value added test scores (see Grosskopf et al. 1997 on how these measures are obtained). The approach of Fukuyama and Weber is an extension of the previous work of Grosskopf et al. (1999) where the authors used an indirect distance function that allowed the measurement of output expansion possible if school districts were able to re-allocate inputs while maintaining a given budget. Banker et al. (2004) also examined efficiency of Texas school districts. They used as inputs three types of expenditures (related to: instruction; administration; and support), and used as outputs the total pupil enrolment in elementary schools, middle schools and high schools.

The foregoing examples relate to the US, where assessments are normally at district level. Elsewhere studies at school level can be found for example in Burney et al. (2013) who look at the efficiency of public schools in Kuwait. Haelermans and De Witte (2012) examined the influence of innovations in the efficiency of Dutch secondary schools. They observed that profiling, pedagogic, process and education chain innovations are significantly related to school efficiency, whereas innovations in the professionalization of teachers are insignificantly related to school efficiency. Portela et al. (2012) also looked at efficiencies of Portuguese schools. We return to this assessment in Sect. 12.2.2.3 as it is linked with an online tool that allows the computation of school efficiency scores online. Another study in Portuguese schools is that in Portela and Camanho (2007) which is a good example of how the two perspectives of assessment, efficiency and effectiveness can be seen as complementary. The authors assessed schools from two perspectives: One called the society perspective (related to effectiveness studies mentioned above), and the other called the educational authorities' perspective (related to efficiency studies mentioned above). If one is evaluating schools from the parents' perspective, the fact that some schools may appear to have low Value Added due to scarce resources, poor location, or poor quality of teachers, is of no particular importance. The objective of parents is just to identify the best schools for their children rather than make allowances for less than satisfactory performance by teachers or schools. However, if one is assessing the schools from the perspective of an authority charged with funding and overseeing the services delivered by schools (the most usual implicit perspective in published papers) all the reasons behind poor or excellent VAs are of interest if steps are to be taken to improve the performance of the schools. In this case an efficiency perspective should be also considered.

12.2.2.2 Using Aggregate Pupil Data to Identify Differential School Effectiveness

It has been found (e.g. Gray et al. 1986; Sammons et al. 1993) that some schools have differential effectiveness depending on pupil prior attainment levels. More generally a school may intentionally or otherwise be using teaching styles which favour more certain groups of pupils compared to others.

Identification of the direction and degree of any differential effectiveness at schools is valuable for a number of reasons. Thanassoulis (1996) argues it helps to identify suitable teaching practices and role model schools for improved all round effectiveness. Where a school streams pupils by ability, appropriate teaching practices and role model schools for raising the level of achievement of pupils of a given ability range can be identified. See also Sammons et al. (1993) on the implications of differential effectiveness for comparing schools. It is preferable to use pupil-level data to ascertain the existence or otherwise of differential school effectiveness as illustrated in the next section.

Using pupil-data, De Witte and Van Klaveren (2014) examine which configuration of teaching activities maximizes student performance. Using data from the Trends in International Mathematics and Science Study, they formulate a nonparametric efficiency model that accounts for self-selection of students and teachers in better schools, and complementary teaching activities. The analysis distinguishes both individual teaching (i.e. a personal teaching style adapted to the individual needs of the student) and collective teaching (i.e. a similar style for all students in a class). Moreover, they test to which group of students the teacher is adapting his/her teaching style. De Witte and Van Klaveren (2014) show that high test scores are associated with teaching styles that emphasise problem solving and homework. In addition, teachers seem to adapt their optimal teaching style to the student representing the 70th percentile on attainment.

In many cases pupil-level data is either not available or not accessible to the analyst. Thanassoulis (1996) puts forth an approach for ascertaining the presence, if any, and direction of differential school effectiveness using aggregate pupil data. The method in Thanassoulis (1996) is based on contrasting schools on the distribution of grades pupils obtain while allowing for the abilities of the pupils, for their family background and for the overall effectiveness in value-added of each school. The method is developed with reference to British secondary schools recruiting pupils at age 11 and measuring their exit achievements at age 16. It can, however, be adapted to other educational systems especially where the only difference is the grading system used and/or the ages of the pupils concerned.

The method in Thanassoulis (1996) begins with an assessment by DEA of the schools concerned on their effectiveness in value added. The set of input variables suggested to assess value added of secondary schools were the mean verbal reasoning score on entry and the percentage of students not receiving free school means (the latter being used as a surrogate for parental background of the pupil). The set of outputs used were the percentage of pupils placed after GCSEs (i.e. in work or further education), and the average exit GCSE score per pupil. GSCE is the General Certificate of Secondary Education which pupils obtained at that time in Britain at the completion of compulsory secondary education, normally at age 16. The number of A, B etc. grades achieved by the pupils of each school were publicly available (and defined here as NAj, NBj, etc. for school j). So it was possible to compute the aggregate GCSE score of school j as $\mathrm{Gj} = 8\mathrm{NAj} + 7\mathrm{NBj} + 6\mathrm{NCj}$

+ 5NDj + 4NEj + 3NFj + 2 NGj + NUj and thereby the mean GCSE score per pupil for each school. (The weights 8 for grade A, 7 for grade B etc. were at the time the accepted approach to converting letter grades to an aggregate numerical GCSE score).

A standard output oriented DEA model was used to compute the effectiveness of each school j_o, whose outputs were the efficiency score (θ^*) and the intensity variables (λ_j^*), the latter denoting how much a peer unit j contributes to the targets of the unit being assessed (j_o).

The grades profile of the efficient comparator "school", cj_o, of school j_o is given by:

$$\begin{aligned} PA_{cjo} &= \sum_{j=1}^{n} \lambda_j^* PA_j \\ PB_{cjo} &= \sum_{j=1}^{n} \lambda_j^* PB_j \\ PC_{cjo} &= \sum_{j=1}^{n} \lambda_j^* PC_j \end{aligned} \tag{12.4}$$

where PA_j is the number of A grades per pupil at school j, (computed as NA_j divided by the number of pupils in the school). $PB_j \ldots$ PU$_\mathrm{j}$ are defined in an analogous manner.

Thanassoulis (1996) suggests that "A comparison of the grades profiles of the efficient comparator cj_o and school j_o can be used to gauge the differential effectiveness of school j_o". The grades profiles of two schools can be compared in a number of ways. Thanassoulis (1996) suggests a simple way where grades A to C (top grades) are used to compute a component "ATOC" and the rest to compute a component "DTOU". The ATOC and DTOU component of school j are:

$$\mathrm{ATOC_j} = 8PA_j + 7PB_j + 6PC_j$$

and

$$\mathrm{DTOU_j} = 5PD_j + 4PE_j + 3PF_j + 2PG_j + PU_j.$$

Thanassoulis (1996) puts forth a procedure for computing the expected E(ATOC) and E(DTOU) components of school cj_o if it had had the same academic effectiveness as school j_o. (The details of how E(ATOC) and E(DTOU) can be computed are beyond the scope of this chapter but can be found in Thanassoulis (1996)). It is then suggested that:

- If ATOCj_o = E(ATOCcj_o) then we have no evidence of differential effectiveness between schools j_o and cj_o;

- If ATOC$j_o \neq$ E(ATOCcj_o) then:
    - if the difference is substantial, we have evidence that schools cj_o and j_o have dissimilar differential effectiveness over pupils of different academic ability;
    - the school with the larger ATOC component is likely to be more effective over the stronger (on entry) pupils.

"School" cj_o would reflect the grade profiles of the efficient peers to school j_o. Thus in effect the method looks for differential effectiveness between school j_o and its efficient peers. Clearly it is difficult to specify a general purpose threshold for the difference in ATOC components which would trigger an identification of dissimilar differential effectiveness between schools as this largely depends on the factors that might have been omitted from the input/output variables used in the DEA assessment. It is clear, however, that the larger the difference in the ATOC components of schools j_o and cj_o the stronger the indication of dissimilarity in differential effectiveness between school j_o and its efficient peers. Thus the method should identify at least the cases where there is substantial dissimilarity in differential effectiveness between schools.

If school j_o turns out to be efficient and self-comparator we cannot draw any conclusions about its differential effectiveness. Further, the comparative basis of the method outlined here means that it may fail to identify differential effectiveness in those cases where the schools being compared have similar differential effectiveness. On the other hand where the schools do differ they become good examples to one another on teaching practices, which can benefit those ability ranges their current teaching practices disadvantage. These difficulties are overcome by attempting to identify differential school effectiveness using pupil level data as outlined later in this chapter.

12.2.2.3 On-line Platforms for Assessment of Schools

Most DEA applications in education, whether using aggregate pupil or pupil-level data, normally rely on analysts, usually from outside the schools concerned, to conduct the assessments of VA. However, there are cases where the assessments are set up in such a manner that schools can self-assess on an on-going basis as new data becomes available. We outline here one such approach, operationalised in Portugal.

The Portuguese Education Ministry supplies every year the media with exam results and they publish school rankings based mostly on these results. Since 2013 the Ministry also provides some contextual variables that allow a contextualized ranking analysis. There is, however, a privately run and innovative platform called BESP (feg.porto.ucp.pt/besp) that provides information to the general public on a set of indicators based on the national exam databases (see Portela et al. 2011). This platform concerns only secondary schools (with students from the 10th year to the 12th year of schooling). It is designed to serve not only the general public, but also to serve schools as a tool for self-evaluation. Indeed, schools can enter data through on-line questionnaires and immediately have access to a number of graphs that show how they are evolving on certain indicators over time, as well as how they perform on specific indicators in relation to a comparable set of schools, which can be in the country, district, or of the same type, etc. A DEA procedure is embedded within BESP, which allows schools to choose a customized set of inputs and outputs and compare their performance with that of other schools.

The information displayed in BESP regarding individual indicators is illustrated in Fig. 12.3.

Graph (a) shows the distribution of the indicator ‘average results on national exam’, with the percentile for the selected school displayed. In the example in Fig. 12.3 the school chosen lies at the 73.39 percentile. In the radar graph (b) of

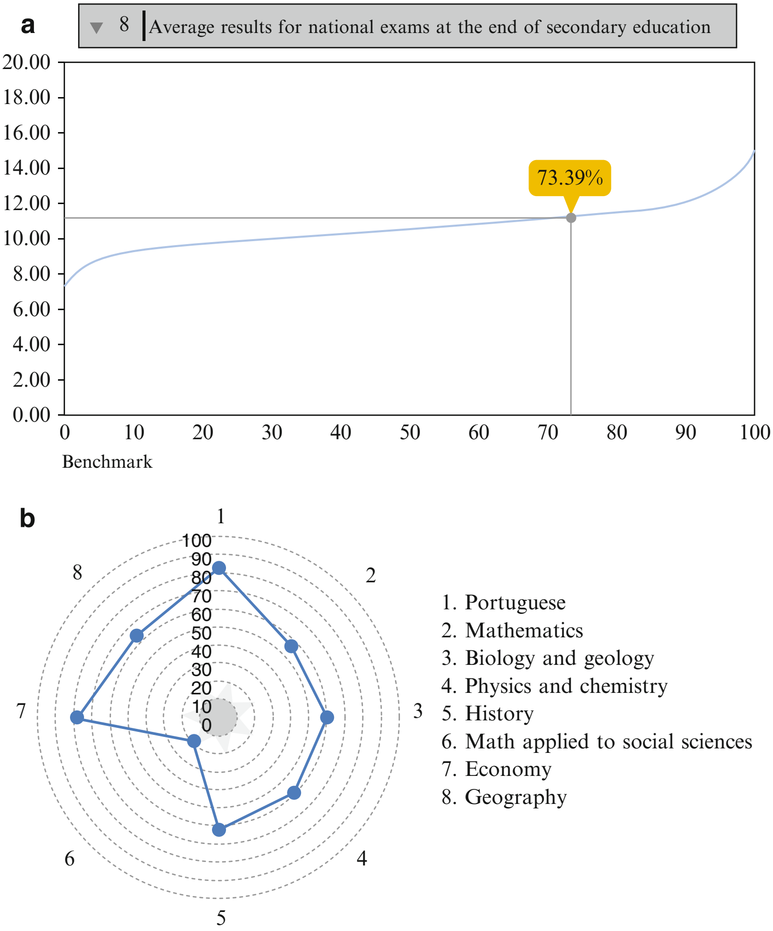

Fig. 12.3 BESP – results for the indicator average classification of the school in final exams (Picture taken from Portela et al. (2011))

Table 12.2 Set of inputs and outputs available within BESP

Inputs	Outputs
Average grades in Portuguese in years t-2 and t-3	Average grades in the kth national exams in year t
Average grades in Mathematics in years t-2 and t-3	Percentage of school students that took exam k
Parents’ average years of schooling	Percentage of students concluding secondary education in the ‘normal’ 3 years
Economic context of the school (computed on the basis of the number of pupils in the school that receive subsidies from the state)	Percentage of students that proceeded to university
	Percentage of students that did not abandon the school

k = Portuguese, mathematics, biology and geology, physics and chemistry, history, economy, geography, mathematics for social sciences

Fig. 12.3 details are shown for the percentiles in which the school lies on each course that is included in the overall average for the school. The radar shows in which subjects the school has better percentile position (e.g. subjects 1 and 7) and the ones where the school has worse percentile position (subject 6).

BESP also enables schools to self-assess using DEA. A menu of potential inputs and outputs (see Table 12.2) are presented to the school from which it can choose a subset on which to be evaluated. The assessment is at the school level using aggregate pupil data.

BESP enables schools to solve on-line DEA models with the inputs and outputs selected from the above list. More details on how to conduct these assessments can be found in Portela et al. (2011, 2012). Given that not many schools upload data into the platform, the assessment possible at the time of writing is the one with the first two inputs and the first output in Table 12.2, for which data are available from public exam databases. Clearly this type of analysis focuses on academic outcomes and neglects other, important non-academic outcomes of education (such as developing inter-personal skills and nurturing responsible social skills).

Focusing on academic outcomes, if a school selects the available inputs and outputs in Table 12.2, the output displayed by BESP is the overall efficiency score and a radar showing how the observed inputs and outputs of the school compare to its potential attainments, controlling for its student intake as reflected in the input variables. In addition, radars showing how the school compares with its peers are also displayed, such that the school can identify the factors where other schools are doing better than itself. An example of such radars is shown in Fig. 12.4 for a school, which has an efficiency score of 79.2 %.

The peers (identified by the darker lines which in general enclose the lighter coloured lines) have similar or slightly higher inputs than the school assessed (the inputs are taken as aggregate grades on entry in this case – ‘EntyGrades’). In spite of that, both peers achieve clearly higher average scores on exit in most of the courses. The above assessment shows the evaluation of the selected school against all schools in the country, but private schools should probably be excluded from the

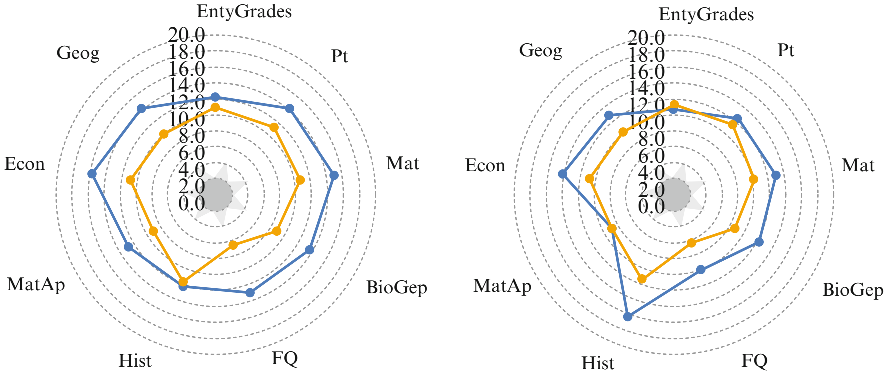

Fig. 12.4 A school being compared with its peers in BESP (Graph from Portela et al. 2011)

comparison set as the selected school is a public school. This can be readily done within BESP.

With the advent of the internet several internet-based tools such as the above have become available to publish information regarding school performance. The publication of results on school performance is normally the responsibility of national governments through their agencies. For example, in the UK, the Department for Children Schools and Families publishes performance tables (education.gov.uk/schools/performance/). The user can select any particular school and gain access to details on the school demographics as well as its performance on a number of indicators, including a measure of value added (VA). Also in the UK there is a web-based application called RAISE “report and analysis for improvement through school self-evaluation” (raiseonline.org), which enables schools to look at performance data in greater depth as part of their self-evaluation process. This application is, however, not available to the general public, but just to schools. This is also the case with the UK School Financial Benchmarking (SFB) website (sfb.teachernet.gov.uk) (for details see Ray 2006; Ray et al. 2009).

In Norway Skoleporten is a national school accountability system, which contains publicly available data on indicators for results, resource use and learning environment (details in Haegeland 2006). The Swedish National Agency for Education also publishes data for all levels of the education system (skolverket.se/sb/d/190). Apart from data on several different indicators, the agency also publishes expected results for each individual school, estimated using linear regression. In Portugal the Education Ministry has a website (infoescolas.mec.pt/) that shows educational statistics for all the schools, including some progress measures similar to VA measures.

In the US there is the Tennessee Value Added Assessment System (TVAAS) developed by Sanders et al. (1997), which has a public website (tn.gov/education/data/TVAAS.shtml) where the general public can access reports on VA. This system has also been adopted by other states, and in particular Pennsylvania, which has the PVAAS, (available at pvaas.sas.com) and Ohio, which has EVAAS (available at ohiova.sas.com).

Note that the trend towards internet-benchmarking platforms that allow on-line and immediate comparisons between production units can also be found in sectors beyond education. For example, the construction industry has a benchmarking platform icBench (icbench.net) (Costa et al. 2007) available in Portugal, whereas in the US there is BM&M (construction-institute.org), and in the UK there is KPIzone (kpizone.com). None of these or the above examples, however, use DEA to carry out performance assessments. An example of a platform that allows also for aggregate assessments through DEA can be found in iDEAs (isye.gatech.edu/ideas) (see Johnson and McGinnis 2011). This platform is targeted at warehouses or other industrial systems and combines benchmarking and DEA to allow managers to benchmark their performance against others. Bogetoft and Nielsen (2005) report an internet-based benchmarking system, applied to Danish commercial and savings banks, that incorporates a DEA model. The developed platform is currently being commercialized and applied to other industries (Ibensoft Aps 2013).

12.2.3 DEA Applications Using Pupil-Level Data

Given the hierarchical structure of education, multilevel models have been the main instrument of analysis in 'educational production function' approaches. To this popularity contributed, amongst others, the development and enhancement of hierarchical modelling techniques, known as **multilevel modelling**, by Goldstein (1987) and Raudenbush and Bryk (1986). Additional impetus was given by Sanders et al. (1997), who carried out a project in Tennessee (USA) which included not only the estimation of the VA of schools but also of teachers. More recent examples of applications using pupil level data can be found in Agasisti et al. (2014), Mancebón et al. (2012) and Hanushek et al. (2013).

There are not very many applications of DEA to pupil-level data. As we have seen, DEA models were initially mainly applied to aggregate (e.g. school level) data. To the authors' knowledge Thanassoulis (1999) is the first DEA study that used pupil-level data, to set achievement targets for pupils. School effectiveness or VA was not investigated in the Thanassoulis (1999) paper. The measurement of VA of schools through DEA, using pupil-level data, was first attempted by Portela and Thanassoulis (2001) and Thanassoulis and Portela (2002). The approach adopted by the authors to compute the VA of schools and the key findings will be detailed in the next two sections.[2]

[2] Note that parametric frontier models have also been widely used in the educational context, but these will not be detailed in this chapter (examples of pupil-level studies through stochastic frontier models can be seen amongst others in Cordero-Ferrera et al. (2011), Deutsch et al. (2013), Perelman and Santín (2011) or Crespo-Cebada et al. (2014)).

12.2.3.1 An Overview of DEA Applied to Pupil-Level Data

We will use the approach of Portela and Thanassoulis (2001) to outline the use of DEA with pupil-level data. Their approach was inspired by the approach outlined in Sect. 12.2.2.1 for disentangling managerial from program efficiency introduced originally by Charnes et al. (1981). The DEA approach mimics the parametric multilevel modelling approach in that pupils are assessed hierarchically so that pupil effects can be isolated from higher level effects such as school effects, local education authority effects etc. The approach is illustrated through Fig. 12.5, where crosses represent students from various schools and the circle-shaded crosses represent students from a particular target school. On the axes we depict attainment of the student on entry to a certain educational stage and attainment on exit from that same educational stage, respectively. These two variables normally feature in VA assessments in secondary education but they are only a subset of the input and output variables generally used. The data in Fig. 12.5 are real, corresponding to grades on entry (at the 9th grade) and grades on exit (at the 12th grade) of students from Portuguese secondary schools in a certain 3-year period. Pupils beyond the global frontier are deemed 'outlier', not permitted to define the frontier. (Two pupils of the target school (shaded crosses) are also deemed outlier.) The data were used within an evaluation programme called AVES (for details see Portela and Camanho 2010).

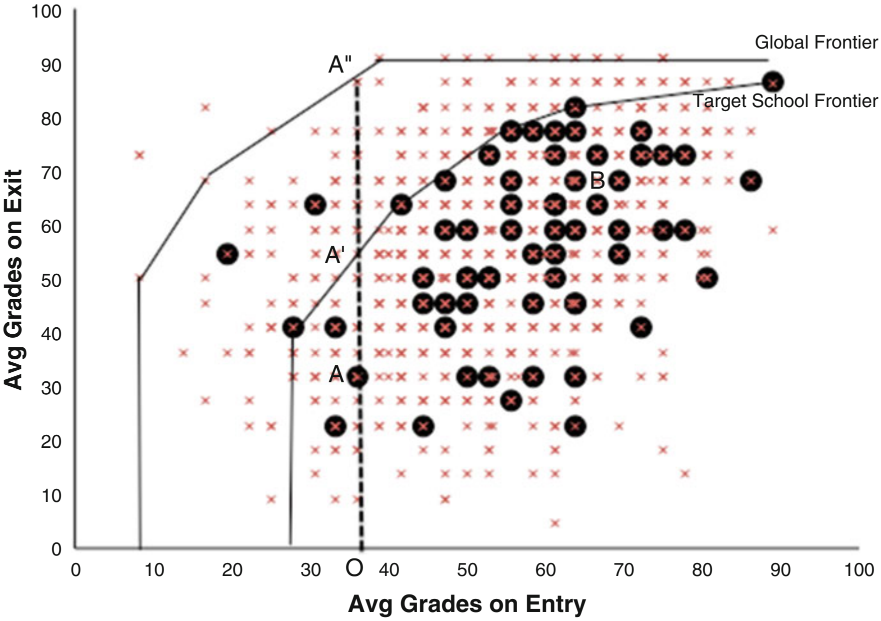

Fig. 12.5 Illustration of the DEA approach to assess VA using pupil-level data

Consider now pupil A attending the target school in Fig. 12.5. The efficiency of this student with reference to the frontier of students attending the same school can be measured through OA/OA′. If instead we compare the student to the overall set of students from all schools (global frontier in Fig. 12.5) the efficiency of student A is measured through the distance OA/OA″. As a result, the global efficiency score (or pupil-within-all-schools efficiency) decomposes into two components:

$$\frac{OA}{OA''} = \frac{OA}{OA'} \times \frac{OA'}{OA''} \tag{12.5}$$

A component that is attributed to the pupil, OA/OA′ (termed pupil-within-school efficiency in Portela and Thanassoulis (2001)), and a component attributable to the school OA′/OA‴ (termed school-within-all-schools efficiency in Portela and Thanassoulis (2001)). The pupil effect (OA/OA′) incorporates the shortfall in achievement of pupil A that is due to the pupil alone, as the target school he/she attends has demonstrated is able to place students with the entry attainments of pupil A at point A′ on exit. The school effect (OA′/OA″’) reflects the component of under attainment of pupil A that cannot be attributed to the pupil (imagining that he/she was at point A′) but to the school, that did not foster as much attainment on exit as other schools in the sample did, for students with entry attainment as that of pupil A.

The pupil-within-school efficiency scores reflect the diversity of achievements attributable to the pupils in that school. Within school efficiencies are not comparable across schools except to reflect relative diversity of attainment of pupils of different schools. For example, in a school where the mean of the pupil-within-school efficiency is lower than in another school the pupils of the former school will in general be further away from their school frontier than in the case of the latter school. However, we cannot say in which school we have higher value added in absolute terms as this will depend on the relative positioning of the school frontiers. In schools where pupils are generally close to maximum possible attainment we expect high average pupil-within-school efficiency scores, whereas a low average is expected for schools with highly heterogeneous pupil attainments.

The school-within-all-schools efficiency reflects the shortfall, if any, between the current best attainment of a student and the corresponding best exit attainment available across all schools for a given entry level.

In Fig. 12.5 we have considered just two levels of analysis, pupil and school. Originally in Portela and Thanassoulis (2001) three levels of analysis were considered the pupil, the school, and the school type (Comprehensive, Grant Maintained and Independent schools). An alternative 3-level analysis is one where pupils are level 1, classes are level 2 (assuming pupils are taught in different classes), and schools are level 3.

Teacher effects on pupil achievement is a recent issue addressed in the literature (see e.g. Chetty et al. 2014), and this effect can be estimated using pupil-level data in the manner outlined here, where a teacher is associated with a given class of pupils. To illustrate the approach to assessing teacher effectiveness let us refer to

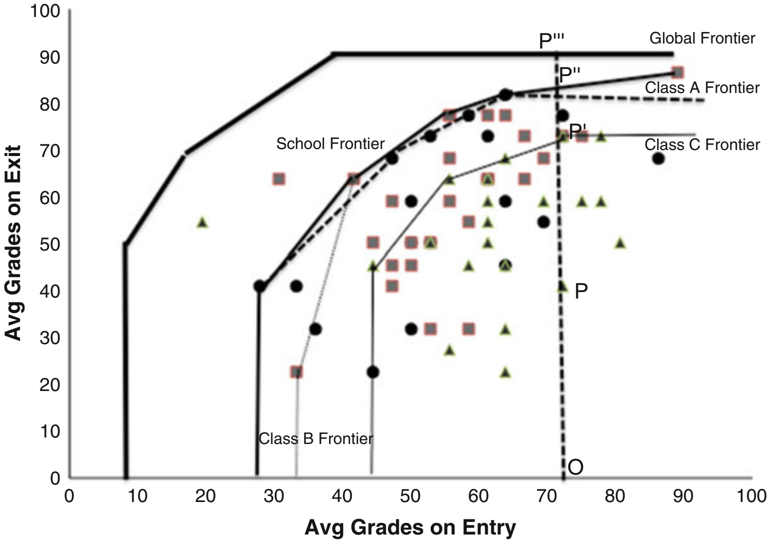

Fig. 12.6 Representing pupils by class within school

Fig. 12.6 where a global frontier has been drawn mapped out by attainments across all schools and their classes. Consider now a target school which has three classes (A, B, and C) for teaching the course under analysis. A frontier for each class can be constructed following the same principles as in Fig. 12.5 and distances computed accordingly. This is illustrated in Fig. 12.6 (where the all-schools (or global) frontier was kept, but students defining that frontier have been omitted).

Frontiers of class A (students from this class are represented by circular dots) and B (students represented by squares) are very close to the school frontier (constituted mainly by pupils from Class A and from Class B), whereas class C frontier (students represented by triangles) lags behind the overall school frontier. For pupil P attending Class C in the school under consideration we have the following decomposition of a global efficiency score (or pupil-within-all-schools efficiency):

$$\frac{OP}{OP'''} = \frac{OP}{OP'} \times \frac{OP'}{OP''} \times \frac{OP''}{OP'''} \tag{12.6}$$

The first term in the right hand side of (12.6) measures the pupil effect, the second term measures the classroom effect, and the third term measures the school effect, on a pupil having the entry attainment of pupil P. If attainment on a single subject is being considered (e.g. maths) the classroom effect is indeed the teacher effect assuming only one teacher teaches the subject concerned. In such a case the component (OP′/OP″) is a measure of pupil under-attainment that is attributable to the teacher of class C as other classes in the same school have shown that, for the initial level of attainment of pupil P, higher attainment than at P′ was possible.

If attainment on entry and on exit cover a varied number of subjects, the component OP′/OP″ will in fact be attributable to the class or more precisely to all teachers teaching the subjects under consideration to the class concerned.

The approach considered here can be generalised to account for other types of effects, depending on the number of categories under which pupils can be grouped beyond classes and schools – e.g. gender, ethnicity, country, etc. For example, country specific effects can be estimated if one assumes that inner frontiers in Fig. 12.6 represent schools, enveloped by a country frontier, and a global frontier enveloping all country frontiers. In fact the approach outlined can be seen as a more general one that can account for categorical variables in DEA assessments, which is originally designed to deal only with continuous variables.

Thieme et al. (2013) have extended the approach using pupil-level data outlined above by considering not only pupil-level variables but also school level variables. Accordingly they divided the school effect into three types: resource endowment effects, peer effects and selection bias effect. The first is related to the fact that pupils with different levels of performance can be placed in schools with different resource endowments; the second is related to positive externalities that can happen when students try to emulate their peers to reinforce their identification with the group; the third is related to the fact that more able and motivated students may place themselves into certain types of schools. The inclusion of school variables in the Thieme et al. (2013) approach works by defining additional frontiers in Fig. 12.5 or Fig. 12.6, where, for example to estimate the resource endowment effect, each pupil would be compared to its own school frontier, to a frontier including all schools that operate with no more resources than the pupil's school, and finally to the global (all-schools) frontier.

12.2.3.2 Applying DEA to Pupil-Level Data

In this section we briefly reflect some of the main practical aspects of using DEA on pupil level data.

Input–Output Variables

The selection of appropriate input and output variables is a fundamental step in any DEA analysis (Thanassoulis 2001, Chap. 5; Emrouznejad and De Witte 2010). The choice of inputs and outputs depends on the perspective from which the assessment is to be carried out. For example, in the case of aggregate level data, as we saw, adopting a perspective of school efficiency led to a different set of input output variables compared to a perspective of school effectiveness. Regarding pupil-level data one is more constrained regarding the inputs and outputs to consider as these may relate to pupils and not to schools (in spite of school variables being also possible to consider). As a result pupil-level analyses use in general two types of variables: categorical and continuous variables. The continuous variables typically include measures of attainment by pupils on entry as an input, and measures of attainment by pupils on exit as an output. Attainments are typically in numerical form e.g. mean grade across a set of subjects or grades by subject (e.g. maths, science, English etc.).

Other continuous variables may relate to age of pupil, or various measures of socio-economic and parental education background (see De Witte and López-Torres 2015 for relevant references). Categorical variables are those whose impact on pupil attainment is to be investigated such as class, school, type of school, gender, ethnicity, etc. As seen before, the most immediate way of considering these types of variables is through the consideration of various frontiers for different categories of the variables (e.g. girls vs males frontiers, school frontiers, classes frontiers, etc.). The literature on the use of categorical variables is also related with the literature on non-discretionary factors (which are in many instances categorical) and initiated with Banker and Morey (1986), and followed by Ruggiero (1998). A note of caution is required when data is subdivided following various categories, as one should assume that these divisions still allow a reasonable number of comparator units to be assessed.

Dealing with Outliers

One of the issues with using pupil-level data is that any chance events (e.g. fluke high or low attainment by a pupil on a subject) is not mitigated in the way it would be when using aggregate data across pupils. As in DEA results are highly influenced by observations which are especially efficient, it is customary to seek to identify 'outliers' on efficiency. In DEA outliers are those observations that have efficiency much higher than 100 % when they are assessed relative to a frontier drawn on all units, excluding the observation being tested for outlier status. Such observations can 'pull' the frontier to largely unattainable levels setting a biased benchmark for other observations to aspire to. Observations with very low efficiency on the other hand, which could be deemed outlier in a statistical sense, do not present a problem in DEA as they do not impact the referent frontier for any units other than the pupil itself (see a discussion in De Witte and Marques 2010).

There are a number of procedures available in the literature for identifying and dealing with outliers in DEA (e.g. see Thanassoulis et al. 2008, Sect. 3.6.4). Earlier methodologies (see De Witte and Marques 2010 for an overview) typically identified and eliminated outlying (super-efficient) data. By removing the outlying observations from the data, one risks losing also the most interesting observations. More recent approaches such as the robust order-m or order-alpha techniques mitigate the influence of outliers without removing them from the sample (see Sect. 12.4.2). It is clear that approaches of this kind for dealing with outliers are very attractive in assessing efficiency in education where outliers at pupil level can be often observed.

Obtaining Efficiency Estimates

The computation of efficiency scores to be included in decomposition (12.5) or (12.6) can be computed through traditional DEA models in the literature. The models typically assume variable returns to scale and are output oriented. The assumption of variable returns to scale is dictated by the fact that attainments on entry and exit are typically percentages or indices not expected to follow a strict mutual proportionality. Output orientation is sensible given that attainment on entry or socio-economic characteristics as inputs, are determined before entry to the stage of education under assessment.

The procedure for computing the efficiency estimates, used in Portela and Thanassoulis (2001) involves the following steps:

- *Assessing the efficiency of each pupil in relation to the global frontier to obtain the pupil-within-all-schools efficiency of the pupil;*
- *Assessing the efficiency of each pupil in relation to its own school frontier to obtain the pupil-within-school efficiency for the pupil.*

The school-within-all-schools efficiency (or VA) at the entry level of each pupil is obtained as the ratio of the pupil-within-all schools to the pupil-within-school efficiency of that pupil. It should be noted that though this approach in Portela and Thanassoulis (2001) is very close to that of program efficiency in Charnes et al. (1981), outlined earlier, it is not identical. The computational procedure of Charnes et al. (1981) if implemented would have implied the use of three rather than two steps. Step (1) would involve computing efficiency scores of pupils within schools as above; step (2) would imply replacing each pupil's outputs by its frontier targets outputs obtained from (1), including non-radial components; Step (3) would require assessing the efficiency of all pupil targets as derived in step (2) to obtain the school-within-all-schools efficiency which is termed 'program' efficiency in Charnes et al. (1981). For larger data sets the approach of Portela and Thanassoulis (2001) is computationally less demanding. The Portela and Thanassoulis (2001) approach captures distances between all frontier points of the school and the all-schools frontiers, whereas the Charnes et al. (1981) approach only uses the efficient part of the school frontier (but not so for the all schools frontier).

Other than classical DEA, which assumes a convex production possibility set, can also be used with pupil-level data. For example, De Witte et al. (2010) modelled the 'production function' as a Free Disposal Hull (FDH) technology. Their paper controlled for outliers using the order-m method of Cazals et al. (2002). See also Thieme et al. (2013) for a similar application.

Aggregation of Pupil Level Results

Although pupil-level scores are of interest *per se*, in pupil-level analyses we are also usually interested in assessing performance at the school or at the school district level. Views about performance at these more aggregate levels can be gained by aggregating in a suitable manner the pupil level efficiencies. Pupil-within-school efficiencies are normally aggregated in the form of a geometric mean. The same can be done for school-within-all-schools efficiencies. When geometric means are used in this way the means are decomposable as indicated earlier into pupil-within-school and school-within-all schools components. This does not hold true if aggregation is by means of arithmetic averages. In both cases, the average (either arithmetic or geometric) reflects the aggregate as long as there is no correlation between efficiency and output, which is pupil attainment on exit in our case (Karagiannis 2015). Otherwise, the aggregation rules proposed by Färe and Zelenyuk (2003) and Färe and Karagiannis (2014) should be applied.

Extracting Additional Information on Performance from Longitudinal Pupil-Level Data

The approach for using pupil-level data in assessing VA at schools as described so far is couched in terms of one period of time only. However, in practice the evolution of VA at a school over time is of utmost interest. In the context of a multi-period assessment of VA, we have an additional consideration in that the global frontier of each hierarchical level (school global frontier) would in general be different for each time period. For example the global frontier enveloping all school frontiers may change from one period to the next, and as a result even for a school whose frontier did not move from one period to the next it may still show say a decrease in VA if the global frontier has improved. One approach to overcoming this problem is proposed by Portela et al. (2013) who suggested the use a stable metafrontier defined by all data for all time periods rather than use an evolving global frontier over time. That is, for an assessment where data is available for the last 5 years, all 5 years of data would be considered in defining the metafrontier. The disadvantage of this approach is that the efficiencies relative to the stable metafrontier will reflect best performance observed over the entire long-term period rather than what might have been attainable only up to some earlier point in time under consideration. The significance of this drawback will depend on the use to which the findings of the analysis are to be put. One can always of course also compute comparative performance up to specific points in time if that is of interest as opposed to a global view of the long-term performance of pupils and schools.

Another drawback of this approach is the fact that as time passes, data from new periods become available, and the replication of the analysis would imply a different metafrontier (with all data observed until the period of analysis). However, in practical examples when a considerable number of time periods are included in the definition of the metafrontier, it is likely that it is approximately stable over time and new data can be added without provoking many changes to the frontier.[3]

[3] In education settings, where the variables used in the analysis are grades obtained in national exams, it is unlikely that many changes happen from year to year, except if the syllabus of the course changes. Therefore, when a reasonable number of time períods is included in constructing a meta-frontier it is unlikely that new time periods will imply big changes in that frontier.

To pursue with the computation of VA change, it is important to note that this can only be done at the school level, as pupils are not the same over different cycles of studies. That is, data usually relate to a cohort of pupils analysed at the end of a certain cycle of studies (accounting for their attainment on entry at that cycle). In the next period of time, the cohort for that same cycle of studies is different, as a new cohort is finishing the cycle. It is possible that the longitudinal analysis could track the same pupils over time and over different cycles of studies. In that case the evolution of the VA at the pupil-level would be possible to assess. However, to the authors' knowledge there is no DEA application focusing on such a longitudinal analysis. A good example of such type of analysis can be seen in the Tennessee value added system.

Portela et al. (2013) define Value-Added Change (VAC) at school level, as an aggregate of the VA scores computed at the entry levels of pupils corresponding to the exit cohorts concerned. Thus denoting VA_s^t the measure of VA of school s for the exiting cohort in period t, and va_{js}^t as a measure of VA of pupil j exiting school s in period t, we can aggregate through a geometric mean pupils' VA values to obtain a measure of the school VA. Thus value added change from period t to t + 1 denoted VAC(t,t + 1) can be defined as shown in (12.7).

$$VAC(t,t+1)=\frac{VA_s^{t+1}}{VA_s^t}=\frac{\left(\prod_j va_{js}^{t+1}\right)^{1/N_s^{t+1}}}{\left(\prod_j va_{js}^t\right)^{1/N_s^t}} \tag{12.7}$$

As mentioned before, VA is a measure of the distance between the school frontier and a referent frontier, and conveys no information regarding how students perform within the school at individual student level. Clearly both the school and the pupil-level measures are of interest for managing value added at a school for the better. Portela et al. (2013) compute a Malmquist index inspired measure to capture performance change over time at pupil-level. The measure is based on measures developed in Camanho and Dyson (2006) and Portela et al. (2011), and is of the following form:

$$M_s=\frac{\left(\prod_j E_p\left(X_j^{t+1},Y_j^{t+1}\right)\right)^{1/N_s^{t+1}}}{\left(\prod_j E_P\left(X_j^t,Y_j^t\right)\right)^{1/N_s^t}} \tag{12.8}$$

where $E_P(X_j^t, Y_j^t)$ is the efficiency score of pupil j observed in period t as computed in relation to the frontier P (the pooled metafrontier).

This index can be decomposed into an efficiency change or catch up component and a frontier shift. As shown in Portela et al. (2013) the frontier shift component is precisely the value added change defined in (12.7), whereas the catch up component (CU) is given by:

$$CU_s = \frac{\left(\prod_j E_{t+1}^s\left(X_j^{t+1}, Y_j^{t+1}\right)\right)^{1/N_s^{t+1}}}{\left(\prod_j E_t^s\left(X_j^t, Y_j^t\right)\right)^{1/N_s^t}} \tag{12.9}$$

That is, the ratio of the average efficiency of pupils in period t + 1 in relation to the school frontier of that period, to the average efficiency of pupils in t in relation to that period's school frontier. It is recalled that CU_s can only convey the relative level of dispersion of efficiencies of pupils relative to the school frontier. That is, a value larger than 1 would imply that pupils in period t + 1 are closer to the frontier of that period than pupils in period t, on average. However no conclusion can be drawn as to whether they perform better or worse in period t + 1 compared to period t as this will depend on the relative positioning of the school frontiers in periods t and t + 1.

12.2.3.3 Putting to Use the Findings from DEA Assessments of Pupil-Level Data

Each DEA application has its own aims and the perspective used varies depending on the stakeholders in whose behalf the assessment is undertaken. We illustrate here some typical uses made of the results from DEA assessments.

Gaining Insights into Components of Performance Attributable to Pupils, Schools or Other Hierarchical Levels

An obvious initial output from the DEA analysis of pupil-level hierarchical data are the various efficiency scores. These are provided at the pupil-level, and therefore their analysis is possible at various levels of hierarchical or categorical aggregation that may apply. Graphs showing for each school the global efficiency scores (pupil-within-all-schools) and the pupil-within-school efficiency scores are of particular interest (Thanassoulis and Portela 2002; Portela and Camanho 2010; Thieme et al. 2013). Such graphs are shown in Fig. 12.7 (using a real data set) for two schools with different profiles.

School A and B show similar pupil-within-school efficiency scores (average of 79 % for school A and 77 % for school B). However, they show remarkable differences regarding the within-all-schools efficiency scores. In school B

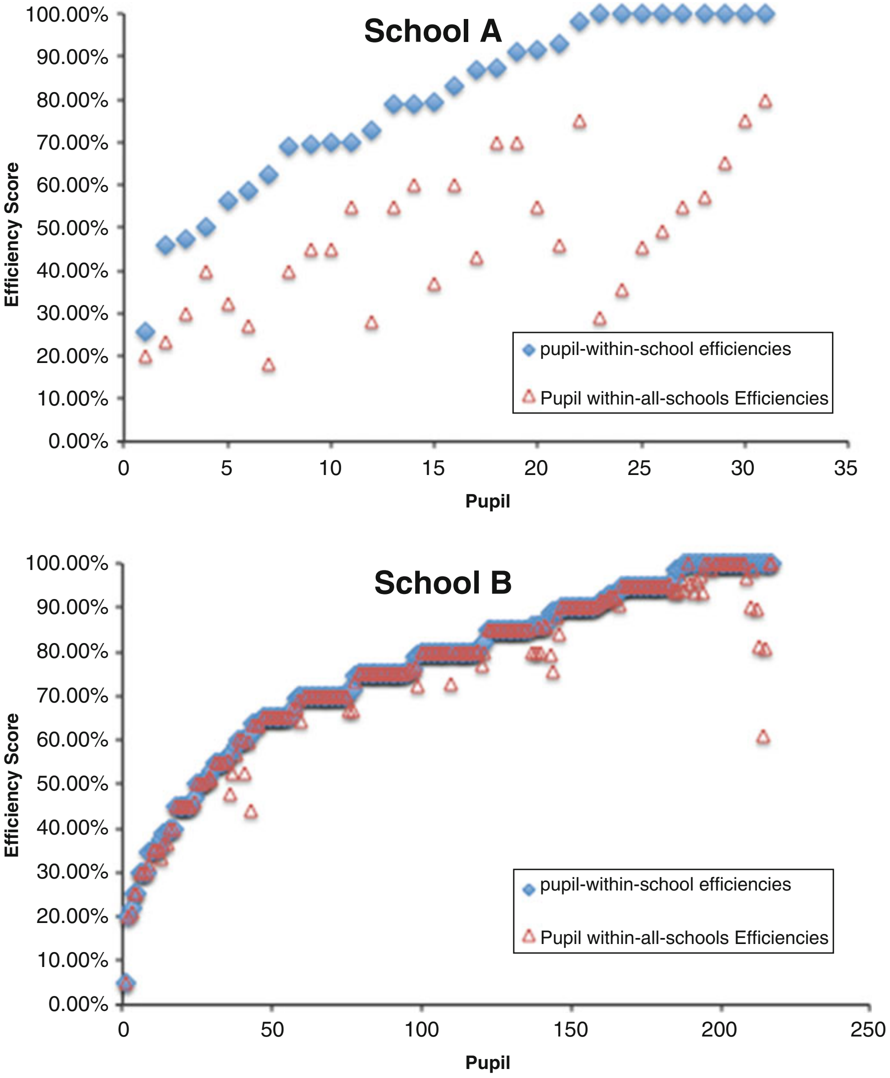

Fig. 12.7 Distribution of pupil-within-school and within-all-schools efficiency scores

pupil-within-all-schools efficiencies are similar to within school efficiencies (75 % on average) and in school A they are dramatically lower (47 % on average). The ratio of the pupil-within-all-schools to the pupil-within-school efficiency, it is recalled, captures the school effect or school VA on a pupil's performance and it is referred to as the school-within-all schools efficiency at the pupil data point. This ratio for school B has an average value of 98 % whereas for school A it has an average value of 60 %. This means that school A has a larger detrimental impact on pupil attainment than does school B.

Information of the foregoing type can be summarised in a tabular form contrasting the average school-within-all-schools efficiency with the average pupil-within-school efficiency for each school. In Portela and Camanho (2010) such tables have been compiled. Similar tables are also found in Portela et al. (2013), where change in school and pupil performance over time are also summarised.

Identifying Role Model Pupils and Setting Achievement Targets

An approach for target setting at pupil-level is detailed in Thanassoulis (1999). The approach is based on the use of DEA with data at two levels – pupil and school – but the approach can be extended to any number of levels. E.g. the classroom, the school, the global set of schools, etc. Targets are based on the projection of the attainments of the pupil to a frontier point, corresponding to the pupil's attainments on entry to the level of education concerned. These targets flow out directly when using DEA (e.g. the DEA software available from www.deasoftware.co.uk). The least challenging for the pupil targets would be those based on his/her within-school class frontier, if data by class have been analysed. Such targets should be attainable by the pupil if his/her teacher were to maintain his/her current level of effectiveness with pupils. Consider pupil P whose attainments to date are not on the efficient frontier. Role model pupils within the pupil P's class, with similar attainments on entry, are identified by the DEA model and their attainments to date can be used to establish the target attainments pupil P should be able to attain. Targets based on higher levels of aggregation of pupils follow a similar logic. For example, targets based on the global all-schools boundary can be set but would be more challenging for the pupil and they could involve a challenge to the school of the pupil too, if the school's within all schools efficiency is not 100 %. Role model pupils may also be harder to hold out as examples to an inefficient pupil if those pupils are from schools other than his/her own.

Except for Thanassoulis (1999) not many studies have explored target setting for pupils. However, personalised target setting is an important part of a school improvement process. Further, the approach also shows clearly the scope for further improvement in pupil attainment that rests on school rather than pupil efforts. When longitudinal data are available the static snapshot results can be complemented with projections on future performance, based on improvements in value added effectiveness in the past. This approach "enables schools and administrators to determine, given the expected growth rate of a particular group of students, what proportion of students will meet a desired standard, and this facilitates planning and resource allocation" (OECD 2008, p. 79). The Tennessee value added system provides projection reports where the trajectory of students is combined with the trajectory of the school's value added. These reports play an important role as "if a large number of students are projected to fall below the proficiency standard, the school has an early warning signal that it must aggressively address the factors… that are retarding student progress" (OECD 2008, p. 80).

Identification of Differential School Effectiveness

As mentioned earlier, early literature identified the issue of differential school effectiveness in the sense that the effectiveness of a school may differ by group of pupils (e.g. by ethnicity, social economic background, innate abilities, gender, etc.) (see e.g. OECD 2008). Differential effectiveness could be identified using aggregate data as we saw earlier, but pupil-level data lends itself much better for this purpose.

Take for example the case of Portuguese schools, assessed in Portela and Camanho (2010). If average VA is computed for each group of students according to ‘innate –ability’, one can have an idea of the profile of the school’s effectiveness with different ability groups. Figure 12.8 illustrates this idea, with two schools, C and D, showing different profiles on their VA for different ability groups (represented on the horizontal-axis).

The VA of School C increases with rising pupil attainment on entry. The profile of this school is very close to that observed across the full set of schools under analysis. However, the school shows in all ability groups lower than average performance. In contrast, school D shows the reverse profile having its highest VA for pupils with the lowest attainment on entry, with VA falling as attainment on entry rises. This means that school D excels at making its weakest pupils on entry attain the best results on exit. This appears to be done at the cost of potentially not raising the attainment of the stronger on entry pupils as far as it might have been possible. Its VA is much below the average for the stronger on entry pupils. Note that in interpreting VA in the above manner care is needed to ensure the group concerned (e.g. of weak or strong pupils on entry) should be of sufficient size for the mean to be representative of the group (see for example Simpson 2005).

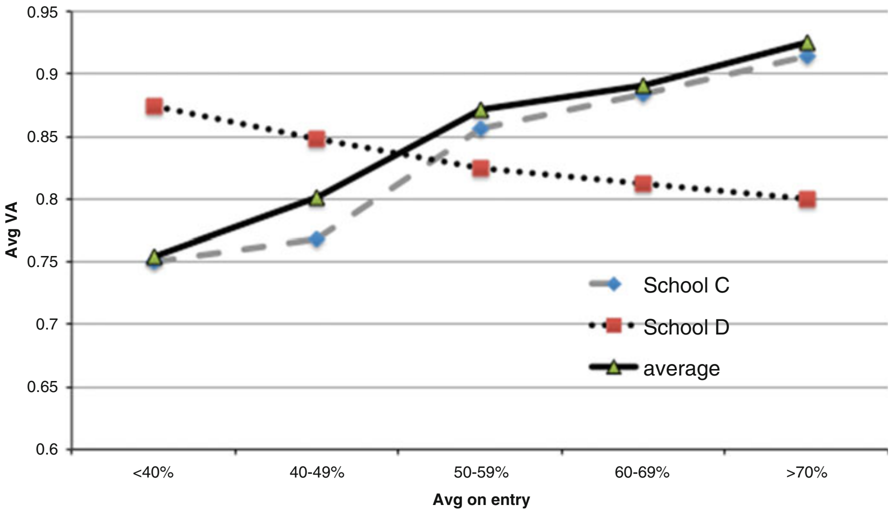

Fig. 12.8 VA of two schools by ability on entry group

The Influence of Environment on School Performance

Estimates of VA are of interest for parents to enable the choice of school for their children. On the other hand policy makers and school administrators have a different interest as it is important to "disentangle the contributions of the school context and the school practice to the gains of the students" (OECD 2008, p. 109). This interest at the policy level can be addressed using pupil-level data in a DEA framework. In particular, second stage procedures outlined earlier for explaining DEA results are a means to this end. There are various types of second stage analysis. One is through qualitative case studies where schools with very high VA and others with very low VA are scrutinised in detail through visits and interviews with school directors, teachers and other staff. For one approach of this type see e.g. Portela and Camanho (2007). Another approach used is a quantitative one where parametric techniques are deployed to regress the VA efficiency estimates on a number of possible explanatory variables. For example, Cherchye et al. (2010) have applied second stage models, but environmental variables were mainly defined at the pupil-level. The exception was the type of school. The type of school was also one of the school level variables used by Mancebón et al. (2012) but in this study several other school characteristics were included in a multilevel model (run in a first stage) to assess the science performance of Spanish students in PISA 2006. Most of the school level variables considered in the DEA assessment were contextual (e.g. proportion of students born in Spain, the proportion of girls in the school, the percentage of students not repeating any grade, average schooling years of mother, etc.). Only two variables concerned school practices: class size and ratio of computers per student. These types of variables are controllable and can be manipulated by policy makers to improve performance of schools. De Witte and Kortelainen (2013) analysed PISA 2006 scores of Dutch students, and performed a second stage analysis where they included a set of variables regarding pupils' background, but also school level variables (from which the contextual variables were the percentage of girls, school size, school autonomy, and average school socio-economic status). They used school practice variables in the form of minutes of maths lessons at the school and student-teacher ratio. Interestingly none of these school practice variables proved relevant in explaining effectiveness of schools. This is in line with previous literature (see the literature review by Hanushek 1986) which found little statistical evidence of the effect of school level variables on the effectiveness of schools.

Choosing Between DEA and Parametric Multilevel Modelling

As discussed in Sect. 12.2.3, multilevel models are the most applied methodology to pupil-level data. Multilevel models are regression-based models, ascertaining the variability in pupil attainment between different levels of aggregation of pupils. With two levels (the most frequent setting, e.g. level 1 is the pupil and level 2 the school) a dependent variable at pupil-level is regressed on a set of independent variables. The intercept is taken as a random variable for each school. Pictorially this is similar to imagining various regression lines, one for each school, where distances between observations and the regression line are taken as the pupil's random error (measuring pupils variability) and distances between school regression lines (or intercepts assuming they are parallel) are interpreted as the school's random error (measuring schools variability, or the school effect).

Some authors have compared DEA based models with multilevel model for evaluating schools effectiveness. For example, De Witte et al. (2010) used robust non-parametric FDH and parametric multilevel modelling on a sample of data relating of 3017 girls attending British single sex schools. They used a robust FDH frontier, and compared the results with those obtained from assessing the same girls using regression-based multilevel modelling. The aim was to contrast the pros and cons of the two approaches. Multilevel modelling is not geared to results at pupil-level. However, it does yield the proportions of variation in pupil attainment attributable to school versus the pupils and these could be compared with similar measures derived from DEA results. It is found the two approaches yield similar results in terms of the proportion of variation that can be attributed to the school and that which can be attributed to the pupil. Correlations found between pupil-schools and pupil-within-all-schools efficiencies are high and above 0.78.

The main disagreement between the two approaches lay on the estimate of school VA where the multilevel model finds a larger proportion of pupil underattainment is attributed to the school compared to DEA. This is related to the fact that the DEA approach looks at how schools compare to each other when the comparison is based on their best attaining pupils (for their entry level) while multilevel modelling in essence captures the school component of pupil attainment by considering average levels of attainment for given entry level. Clearly both perspectives have their own advantages and disadvantages. Methods based on averages have the problem of sensitivity to exceptionally well or poorly performing pupils (for their entry level). That is, a cohort with some students performing well above or for that matter well below expectation may have undue influence on mean values. On the other hand, the comparison between frontiers as in DEA can be less prone to exceptionally good or poor performance provided care is taken to exclude prior to the DEA assessment outlier observations as discussed earlier. This, however, could be underestimating school impacts across the full body of pupils if the school has differential effectiveness (see Sect. 12.2.2.2) and only facilitates best performance for a small subset of pupils.

Given the fact that the information provided by the two methods is based on alternative (average versus 'efficient') views of performance, DEA and multilevel modelling should be seen as complementary rather than alternative methods (De Witte et al. 2010). This is also the conclusion arrived at in Portela and Camanho (2010), when looking at the differences between the application of the above DEA approach and the approach that at the time was being applied by the UK Department for Children School and Families in constructing public league tables on the VA of schools. Other studies have applied both methodologies, but not necessarily to compare them. For example Mancebón et al. (2012) used the two methods in a complementary manner, where multilevel models were used to identify the underlying production function, and then a DEA model was constructed, based on the evidence of the relevant variables in the multilevel model, to compare schools. The interest was mainly in a comparison between publicly funded and fee charging schools.

12.3 Applications of DEA in Higher Education: Institution Level

Education at any level involves the augmentation of human capital. But there are some key differences between the various levels of education. Primary education is typically delivered by generalist teachers, while secondary education involves specialist teachers. Nevertheless, at least at lower secondary level, the curriculum is broad and this imposes a degree of similarity of experience of students at different schools. At higher education level, the experience is considerably more specialised, and the distribution of subject specialisms varies more across institutions.

There are some additional features of higher education that serve to distinguish it from primary and secondary schooling, and these might make the sector intensely competitive and hence have implications for efficiency. For example, in most countries education is compulsory for students up to some level of secondary, but participation in higher education is optional. In addition, primary and secondary education is typically financed through the tax system, but students in higher education in many countries have to pay (often substantial) tuition fees. Finally, higher education is often undertaken during the early years of adulthood, and so students are no longer geographically constrained: many students choose to study at a location that is distant from their parental home.

Unlike staff in secondary or primary education, most if not all academics employed in higher education institutions (HEIs), are contractually expected to undertake research. Hence specialist teaching represents only some of the output of HEIs. Within each subject area, institutions also produce research and engage in knowledge transfer activities – that is, they first create new knowledge, and then they work with various organisations to exploit it.

All of the above considerations throw into sharp focus the need to consider, in any evaluation of higher education efficiency, the nature of providers in this sector as multi-output organisations. This, alongside the availability of good quality data, has made the higher education sector an important test bed for empirical applications of frontier methods generally and DEA in particular. DEA has been employed to assess cost efficiency and technical efficiency. Network DEA can offer insights into the production process and provide more detail to managers and policy-makers on how to improve efficiency.

Table 12.3 Set of inputs and outputs used in Thanassoulis et al. (2011)

Input	Outputs
Total operating cost (measured at constant prices and including depreciation, but excluding catering and student accommodation)	FTE undergraduate numbers in medicine
	FTE undergraduate numbers in sciences
	FTE undergraduate numbers in non-sciences
	FTE postgraduate numbers
	Research funding (quality-related funding council grants and other research grants, measured at constant prices)
	'Third mission' or 'knowledge transfer' activity, measured by income from other services rendered (again, at constant prices)

Note: FTE represents full-time equivalent

12.3.1 Assessments of Cost Efficiency in Higher Education Institutions Using DEA

Assessments of cost efficiency provide potentially useful information on economies of scale and scope and therefore can inform decisions on, for example, how the higher education sector might best be expanded, as well as on how HEIs might become more cost efficient.

An analysis of costs and efficiency in English higher education institutions is provided by Thanassoulis et al. (2011).[4] This analysis includes an input-oriented variable returns to scale DEA model using 3 years of data (2000–01 through 2002–3). The input-output variables are shown in Table 12.3.

Data are sourced from the Higher Education Statistics Agency and cover some 121 English HEIs. These institutions vary considerably in nature – from the ancient universities of Oxford and Cambridge through to civic universities and the new universities of the 1960s to institutions that received university status after 1992. In some institutions costs are distorted by the presence of substantial medical facilities, while in others they are not. The analysis is therefore undertaken both by considering the sample as a whole and separately within four distinct groups:

– pre-1992 institutions with medicine;
– pre-1992 institutions without medicine;
– institutions that were granted university status around 1992 (typically former polytechnics); and
– institutions that have gained university status more recently (typically former colleges of higher education, often affiliated with the GuildHE mission group).

[4] This follows earlier work by, for example, Athanassopoulos and Shale (1997) and Johnes (1998, 1999).

In conducting the DEA, a small number of outliers was identified and excluded from subsequent analysis.[5] The mean DEA technical efficiency of the sample (pooled over time and type of institutions but excluding outliers) was estimated to be about 86 %. This figure appears to be quite high, reflecting the competitive pressures that exist in this sector. The observation at the first quartile had an efficiency score of 79.3 %, again suggesting that inefficiencies are largely competed away. There was, nonetheless, a tail – the minimum efficiency observed was 27.5 %. This likely reflects heterogeneity in the sample of institutions. In particular, small, specialist institutions are likely to have costs that are high in relation to their outputs. This underlines the importance of conducting a more disaggregated analysis.

DEA is applied separately to institutions in each of the four subgroups identified above. Given that this essentially involves estimating a separate frontier for each group, comparisons across groups are valid only in terms of measures of relative homogeneity of efficiency. Mean efficiencies varied considerably across groups and was lowest for institutions that have gained university status more recently. Within this last group there was a tail of institutions where efficiency was below 0.3. This likely reflects heterogeneity, since this group comprises a particularly wide diversity of institutions – from specialist agricultural and arts colleges to generalist universities. Results for this category of institutions should therefore be treated with an appropriate degree of caution.

The DEA model assumed variable returns to scale. Institutions that have increasing, constant, or decreasing returns to scale were identified using the DEA models applied to separate sub-groups of institutions. Most HEIs have constant or decreasing returns to scale. This gave HEIs further information as to how they might be able to change scale size in order to exploit economies of scale. We return to this point below.

In order to estimate an 'efficient' unit cost for each type of output one of the approaches proposed in Thanassoulis (1996) termed 'DEA-RA' (DEA followed by Regression Analysis) was deployed. The purpose of this approach is to obtain estimated parameters for the efficiency frontier. This is done as follows: DEA is performed to identify the efficient and inefficient HEIs. Inefficient HEIs are then projected on to the Pareto efficient frontier by estimating efficient output levels for them using an output oriented radial DEA model, adding any slacks to radial projections. Finally, total operating cost (in pounds sterling) was regressed against the vector of 'efficient' output levels to derive an estimated linear cost function. For the full sample of institutions, the cost equation estimated was given by

$$\begin{array}{ccccccc} C = 13121X_m & + 5657X_s & + 4638X_a & + 3829X_p & + 1376R & + 1537K \\ (12.0) & (19.6) & (18.9) & (7.1) & (84.9) & (14.4) \end{array} \tag{12.10}$$

[5] Following Thanassoulis (1999), the outliers have super-efficiencies (Andersen and Petersen 1993) in excess of 100 % and there is at least a 10 percentage point gap between the super-efficiencies of the outliers and the other observations.

where t-statistics appear in parentheses.[6] Here C denotes costs, X_m, X_s, and X_a denote respectively the output of undergraduates in medicine, sciences, and non-science disciplines, X_p is the output of postgraduates, and R and K denote respectively research income and knowledge transfer as defined in the outputs above. This specification of the cost function assumes no fixed costs. Thus in effect the estimated expression is a linear approximation to the CRS part of the piecewise linear VRS frontier.

The unit output costs derived are reasonable and in considerable accord with the unit costs of the same outputs estimated by a quadratic cost function using the same data in Johnes et al. (2005); they indicate that, at undergraduate level, medical education is the most costly to provide at just over £13,000 (or about US$20,000) per student. This is followed by tuition in the other sciences. Postgraduate education at £3829 per student is estimated to be, on average, less costly to provide than undergraduate education. DEA is not likely to have a very accurate picture here as, more than at UG level, institutions offer very diverse postgraduate courses ranging from expensive MBA degrees to much cheaper PhD degrees. The latter are likely to lower the estimated cost per postgraduate student further because many research postgraduates engage in both undergraduate teaching and joint research activity with staff; while the one-to-one supervision that such postgraduates receive is resource-intensive, their activities also serve to reduce the costs associated with undergraduate provision and with research.

Similar equations showing the relationship at the efficiency frontier between costs and outputs are derived by Thanassoulis et al. (2011) for each of the groups of institutions defined above. The broad picture, with medical education being the most costly, followed by other sciences, is replicated across all groups. The costs associated with postgraduate education vary markedly across different types of institution, however, this being most costly in pre-1992 institutions without medical schools.

Broadly comparable results are reported using a suite of alternative, parametric, estimation strategies, using as data the full sample of institutions. These include stochastic frontier, random effects, and generalised estimating equations to evaluate the parameters of a quadratic cost function, from which average incremental costs associated with each output type are calculated (Baumol et al. 1982). Whatever method is used, the cost associated with undergraduate tuition is highest for medicine, followed by the other sciences. In each of these statistical methods, however, the average incremental costs associated with postgraduate provision is estimated as being higher than that associated with undergraduate provision (other than in medicine). DEA is likely to give a better estimate of postgraduate cost per student once the sample is subdivided into four types of HEI as postgraduate provision is now more homogeneous within each sub-group.

[6] The t statistics should be treated with caution. They are high because the regression fits a line through a scatterplot that comprises observations that lie perfectly on piecewise linear segments.

By pooling the data over the 3 years, Thanassoulis et al. (2011) also investigate change over time in both the frontier and the position of each institution relative to that frontier. This is done using the Malmquist index approach.[7] It is established that, over the period from 2000–1 through 2002–3, the Malmquist index (measuring total factor productivity) shifted very little for pre-1992 universities without medical schools and for post-1992 universities. This index declined quite markedly in the other two groups, however, the median institution suffering a 6 % drop in productivity. Decomposing this change into the components due to shift of the frontier and changes in efficiency of individual units indicates that the decline is *all* due to a shifting frontier. The authors note that this may be an artefact of the data. To be specific, over this period prices associated with the purchases of higher education institutions tended to be rising more quickly than is indicated by general price inflation; consequently the data used for real operating costs may overestimate the real value of inputs in the later years of the study. It is not clear, however, why this would affect some types of university but not others.

Another aspect investigated by Thanassoulis et al. (2011) was the possible augmentation of output levels, notably student numbers, that would be feasible at current levels of expenditure if inefficiencies were to be eliminated. They did this using the output oriented DEA model in two ways. Firstly the model was used in its classical format which scales all outputs equiproportionately maintaining the mix of all outputs (students, research and third mission) in order to gain Pareto efficiency. The potential output augmentations based on this model showed that across the sector there was scope for about 10 % rise in undergraduate science, 15 % in non-science undergraduates and 17 % in postgraduate student numbers. About two thirds of these gains were possible through the elimination of technical inefficiency and the remainder through the additional elimination of scale inefficiencies (i.e. exploiting economies of scale). Looking at the different types of institution the largest rise in student numbers possible in relative terms was at higher education colleges ranging from 20 % for undergraduate science to 36 % for postgraduate students through a combination of scale and technical efficiency gains.

A second variant of the DEA model that Thanassoulis et al. (2011) used involved varying the priorities for output expansion so that only student number augmentations are used to gain efficiency. There were significant differences between these and the preceding results when priorities were uniform across all outputs. They report that when both technical and scale inefficiencies had been eliminated the percentage rise in science undergraduates doubled from 11 % to 22 % and there was a 10 percentage point rise in the number of postgraduate students from 17.52 % to 27.16 %. The least change was in undergraduate non-science students where the percentage gain rose from 15.26 % to 19.81 %. These were large potential gains because the model is such that it seeks for each HEI to raise those student numbers

[7] The index developed by Malmquist (1953) was adapted for use in a DEA context by several researchers in the 1990s. See, for example, Førsund (1993) and Färe and Grosskopf (1996a).

where the maximum gain in absolute terms can be made, unconstrained by the need to maintain the mix of outputs. In some cases the model suggested only one type of student be augmented (e.g. at one university the numbers only of science students rise), because that is where the maximum potential for gain in student numbers lies within given resource levels. In this sense the results represent the potential for gains not only by eliminating scale and technical inefficiency, but also eliminating '**allocative**' inefficiency in the sense of maximising aggregate student numbers by altering the mix of students where appropriate. The authors do, however, sound a note of caution as the model may be overestimating potential gains as the four categories of students used are not sufficiently uniform within each category and so DEA by its nature would base results on those institutions which have the 'cheapest' type of student within each category (e.g. there may be a substantial cost differential between educating say mathematics and biology students yet the model treats both types as simply science students).

The data set used by Thanassoulis et al. (2011) has been used to derive further results by Johnes et al. (2008). This work focuses on statistical approaches and includes consideration of stochastic frontier methods that allow evaluation of efficiency scores while estimating parametric cost functions. This work is usefully considered alongside non-parametric approaches such as DEA. The statistical approach, pioneered by Aigner et al. (1977), has, like DEA, its origins in the work of Farrell (1957), but rather than using linear programming to find the frontier it employs a variant of regression analysis in which the unexplained residual term is defined to include a non-normally distributed component due to inefficiency. By taking this approach, the full toolkit of statistical inference becomes available. The results obtained by Johnes et al. (2008) are broadly in line with those produced by DEA and discussed earlier – efficiency scores obtained using the different methods are positively correlated (though the correlation is not particularly strong). The study is notable for its attempts to include location and the quality of student intake as determinants of costs, though neither appears to be statistically significant.

Johnes and Johnes (2013) provide an update of these statistical frontier analyses, and, using panel data, allow for heterogeneity across institutions by using the latent class variant of the stochastic frontier model (Lazarsfeld and Henry 1968; Orea and Kumbhakar 2004; Greene 2005). The results are broadly supportive of earlier studies.[8] A more refined method that can be used to accommodate heterogeneity is the random parameter stochastic frontier model (Tsionas 2002; Greene 2005), and this is used in another study by Johnes and Johnes (2009). Once again, the qualitative nature of the results confirms the findings of other studies. Broadly

[8] We should note, however, that, when the panel is broken into several sub-periods and models estimated on each sub-period separately, the magnitude of some parameters varies widely across sub-periods suggesting that the results should be treated with caution. Moreover, the latent classes determined by the data are puzzling: one might expect a priori that each class would comprise HEIs with common characteristics (perhaps with research intensive institutions, and other institutions in another). But this is not the case, and the common factor relating the HEIs in a group is not obvious.

speaking, as more allowance is made for inter-institutional heterogeneity, the efficiency score attached to the typical institution increases, though outliers at the bottom end remain.

This raises an important conceptual issue surrounding the evaluation of efficiency. Some institutions produce a given vector of outputs at a higher cost than other institutions for quite legitimate reasons. For example, the ancient universities have real estate that is expensive to maintain and that may be less than ideally suited for purpose; their costs are therefore high relative to those of other institutions. This should not be considered a reflection of inefficiency, as these universities are providing a wider service to society through the maintenance of architectural heritage. Now there may be any number of factors of this kind that explain higher costs in one institution than another. Whether any one of these factors is legitimate or not – and hence whether the higher costs are due to inefficiency or not – is essentially a judgement call. While DEA and other frontier methods produce output that may be interpreted as measures of efficiency, there is always scope for debate about what exactly this output means.

12.3.2 Assessment of Technical Efficiency in Higher Education Institutions Using DEA

The cost function approach of the previous section assumes that firms wish to minimise costs (a potentially dubious assumption in the context of a not-for-profit sector such as higher education). Technical efficiency provides an indication of how well (efficiently) HEIs are using their physical inputs to produce outputs. DEA can be used to estimate output distance functions and hence technical efficiency in this context. While most of the studies which examine technical efficiency are at the level of the HEI, DEA can also be applied equally to data at student level. This compares with the assessment of secondary schools using pupil-level data as described in Sect. 12.2.3. Such student-level studies can be useful in disentangling the effects of HEI efficiency from that of a student's effectiveness (Johnes 2006b, c). This type of information is useful for choosing a strategy for improving both institutional and student value added.

Johnes (2008) provides an example of an output distance function for higher education estimated using DEA.[9] Staff (both academic and administrative), students (both undergraduate and postgraduate) and expenditure on academic services are the inputs into the process which produces teaching (graduates from

[9] Earlier studies using DEA to estimate output distance functions for higher education include Athanassopoulos and Shale (1997), Flegg et al. (2004) and Johnes (2006a). The last is noteworthy for its pioneering application of statistical tests for comparing nested DEA models (Pastor et al. 2002) and for testing for differences in production frontiers of distinct groups of DMUs (Charnes et al. 1981).

undergraduate and postgraduate programmes) and research (income for research purposes). A Malmquist index productivity analysis finds that productivity has grown (on average) by 1 % per annum over the period 1996/96 to 2004/05 and that this is a consequence of improvements in technology that have outweighed decreases in technical efficiency. Rapid changes in the higher education sector over the study period (such as growth in student numbers and the use of online support materials for example, routine use of online multiple choice questions and virtual learning environments) appear to have had a positive effect on the technology of production (pushing the frontier outwards) but this has been achieved at the expense of lower technical efficiency (as inefficient HEIs have struggled to keep up with best-practice performance).

One problem with these results is that they are based on a set of inputs and outputs which do not incorporate quality of student intake and of exit qualifications. A more recent study which attempts to address this problem (at least in terms of undergraduate teaching inputs and outputs) focuses on the effects on efficiency of mergers (Johnes 2014). Undergraduate student numbers are adjusted by entry qualification while graduates from undergraduate programmes are adjusted by category of degree result. The remaining inputs and outputs are as in the earlier study. An output-oriented DEA is applied to an unbalanced panel data set from 1996/97 to 2008/09. The sample is unbalanced for a number of reasons. First, some HEIs merged during the study period. Following merger the new institution was treated as a different entity from the HEIs which merged to form it. In addition, some HEIs entered the data base[10] during the period.

The results of applying DEA to the pooled data set indicate that technical efficiency across the sector is around 80 % (similar to estimates of cost efficiency). The study also makes a preliminary examination of the effect on efficiency of merger activity. HEIs are identified as pre-merging (those institutions which will merge at some stage in the study period), post-merger (those institutions formed from unions of others) and non-merging. The DEA results suggest that post-merger HEIs are typically more efficient than either pre- or non-merging HEIs. These broad conclusions are confirmed using parametric techniques. It is worthy of note, however, that the underlying characteristics of pre-, post- and non-merging HEIs are very different and so the observed efficiency differences could be a consequence of something other than merger. Moreover, a closer examination of the individual mergers indicates that while *mean* efficiency is higher following merger, the efficiency effects can vary by case and there are both winners and losers in the merging process.

Some recent work on efficiency in higher education has focused upon international comparisons. Agasisti and Johnes (2009), for example, use (both constant and output-oriented variable returns to scale) DEA models to compare the performance of institutions in Italy and England over the period between 2002–3 and 2004–5. This analysis employs a rich set of input variables, with data on the student

[10] Data were obtained from the Higher Education Statistics Agency (HESA).

intake, staff, and financial resources; as outputs, numbers of graduates at various levels and a measure of research activity are used. The analysis is conducted both by running separate DEA exercises for the two countries and – as a distinct exercise – running a DEA on the data combined across countries. From the latter analysis, it is established that technical efficiency measures are typically lower in Italian institutions than in their English counterparts; the mean technical efficiency for Italian institutions is just 64 %, compared with a mean score of 81 % in England. In the country-specific analyses, the mean efficiency of institutions is virtually identical in England and Italy, suggesting that the efficiency differences observed across the two countries are primarily attributable to country level effects. Meanwhile analysis of the Malmquist indices, suggests that the Italian institutions are closing the gap. While little change in total factor productivity is observed in English institutions over this period, average efficiency of Italian institutions increased.[11] This finding is in line with the characteristic catching up process whereby less efficient institutions learn good practice from their peers.

12.3.3 Assessment of Research Performance of Higher Education Institutions Using DEA

12.3.3.1 Identifying and Measuring Inputs and Outputs

Research activity may be viewed as a multi-input, multi-output production process with execution time that notably differs across disciplines and even in fields within disciplines. It involves several forms of human (e.g., academic staff, PhD students, research assistants), tangible (e.g., scientific instruments, materials) and intangible (e.g., accumulated knowledge, social networks) resources that are combined to produce an output called “new knowledge”, which has also tangible (e.g., publications, patents, conference presentations) and intangible (e.g., tacit knowledge, consulting services) features. Besides these, research output has two other aspects that are of special interest in assessment exercises: quality (i.e., research excellence) and a value or impact, with the latter being measured by citations counts when academic impact is concerned and judgmentally when impact is in non academic (e.g. business or government) domains (e.g. in the UK Research Excellence Framework of 2013 non academic impact of research was given a weight of 20 % (http://www.thinkwrite.biz/pdfs/quick_impact.pdf) compared to 65 % for conventional research output such as journal articles). The decision-making units behind research activity vary depending on the level of analysis from individual researchers to institutions, such as departments, schools, or even the university as a whole.

[11] The productivity of institutions on the frontier in Italy slipped back over this time period, but the gain in efficiency of other institutions more than compensated for this, yielding an average efficiency increase across the country of a little under 10 %.

The foregoing make research assessment exercises quite a complicated task requiring several assumptions and simplifications to be made at the outset. The first of them concerns the length of the assessment period considered. This is clearly related to the length of the publication period. Both the period from a paper's date of submission to a journal and its acceptance and the period from acceptance to actual publication date differ even within the same discipline. This is due to among other factors the procedures followed by different journals (e.g., number of referees, review rounds, etc.). The shorter the assessment period considered the higher the likelihood that research performance measures will be affected by random factors. This is particularly true for evaluation exercises conducted at the individual researcher level and less for more aggregate levels of analysis, i.e., departments, schools, or universities. Even though there are no a priori norms, empirical evidence from bibliometric studies (Abramo et al. 2012c) suggest that the preferable assessment period is between 3 and 5 years, depending upon the academic discipline considered.

On the input side, measurement of production factors other than labour is in most of the cases difficult or even impossible due to lack of data. We thus usually assume that resources (i.e., scientific instruments, materials, etc.) available to the evaluated units are the same at least within the same field and within a given institution. In addition, we assume, unless data are available, that the hours available for research are the same for each individual in a given field category. This is a reasonable assumption for higher education systems where hours devoted to teaching are established by national regulations and are the same for all, regardless of academic rank. In this case, research can be evaluated separately from teaching as labour input is allocated between research and teaching in fixed proportions. It is a less reasonable assumption for higher education systems where there is a trade-off between research, teaching, and administrative tasks and this should explicitly be taken into account in the assessment exercise.

However, the cost of time is different and is reflected in the labour cost that varies across academic ranks. Since salaries of full, associate and assistant professors differ it would be appropriate to distinguish between them by including three different "types" of labour in the assessment exercise or to measure labour input by its cost if information on individual salaries is available. The purpose of this is to distinguish between different degrees of quality among the employed human resources. Using uniform labour input instead of labour cost will normally have a more severe impact at the individual researcher rather than at more aggregate (i.e., department, school or university) level.[12] For evaluation studies at the individual researcher level when information on salaries is not available or salaries

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

are equal within academic ranks as in some higher education systems (e.g., Italy, Greece), the second best option is to evaluate research productivity by academic rank (Abramo et al. 2013a). Nevertheless, empirical evidence from a recent bibliometric study by Abramo et al. (2010b) indicate that the effect of switching from uniform labour input to cost of labour seems to be minimal expect for outliers.

On the output side, as the intangible counterpart of research output is hard to measure, we consider only codified new knowledge in assessment exercises. These include articles in academic journals, research monographs, patents awards, and presentations in conferences, and their relative importance that differs by subject category and/or discipline. The most prevalent form of codification for research output is publications in academic journals, which is considered as an acceptable approximation of research output in many fields but less so in the arts, humanities and a good part of social sciences (Abramo et al. 2014). But as patents are often followed by publications that describe their content in the academic area and conference presentations usually precede publication of academic work, consideration of the number of publications alone to approximate research output may actually avoid in many cases a potential double counting. Publications may be further distinguished by the type of outlet where they are published into academic journals, chapters in edited books or proceedings, research monographs, reports, theses, etc.

The way we count publications may induce biases in performance evaluation as the number of co-authors as well as the quality/prestige of the publication outlet are factors that one may want to control for in measuring research output.[13] Following Lin et al. (2013) and Hagen (2014)[14] there are four counting methods for collaborative papers: whole counting, where each collaborating author receives full credit; straight counting, where the most prominent collaborator (being either the first author or the corresponding author) receives full credit and the rest receive none; fractional counting, where credit is shared either equally by the collaborators (simple fractional measure) or based on some predetermined weights (full fractional measure); (Simple fractional counting is appropriate when authors are listed in alphabetical order or when it is explicitly stated that authorship is equally shared. Full fractional counting may be based on weights provided by field experts or some other authority.) harmonic counting, where credit is determined by the formula $\frac{\left(\frac{1}{\mathrm{i}}\right)}{\left(1+\left(\frac{1}{2}\right)+\left(\frac{1}{3}\right)+\ldots+\left(\frac{1}{\mathrm{N}}\right)\right)}$ with *i* being the position of an author in the by line and

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

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

N the number of collaborators.[15] For arguments in favour of or against each counting method see Hagen (2014), Lin et al. (2013) or (Abramo et al. 2013b).[16]

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

Turning to research output quality two alternative measures have been proposed in the literature: the journal's impact factor and articles citation counts. Even though many will argue in favour of the latter, the reliability of citation counts in reflecting the quality of an academic article depends on the time lapse between the publication date and the timing of observing the number of citations received. Citations observed at a point in time too close to the date of publication will not necessarily offer a quality proxy that is preferable to impact factor. According to Abramo et al. (2010a), if we do not have data on citations counts for at least a period of 2–5 years (depending on the academic field) after the end of the evaluation period considered it will be preferable to use journal impact factor to approximate publication quality. Nevertheless, since the distribution of both citations and journal impact factors are typically skewed to the right in all academic fields it seems appropriate to use the percentile as means of standardization (Abramo et al. 2010a).[17]

When there are data for a sufficient period of time after the end of the evaluation period considered, citations counts can be used not only as a quality ladder to adjust publication counts but also as an additional research output metric accounting for the value of academic achievements. Academic publications embedding new knowledge have different values measured by their impact on academic achievements. Citations represent a proxy measure of the value of research output that is

[15] For example, for a two-author paper, the first author receives 2/3 and the second 1/3 of credit. For a paper where three authors are involved, the first author receives 6/11, the second 3/11 and the third 2/11 of credit.

[16] There is also a disagreement on whether the choice of counting method affects more papers or citations counts. Lin et al. (2013) found that it impacts citation counts more than paper counts while Abramo et al. (2013b) reached the opposite conclusion.

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

usually included in assessment exercises, in spite of limitations due to negative citations and network citations. Note that when counting citations different weights may be given depending upon the citing article influence, the journal in which it is published, etc.

In order to make citation counts a meaningful metric of research value, their total number should be standardized especially for comparisons across fields to reflect differences in citation intensity as well as the various degrees of covering of each academic field in the existing citation databases. This will render citation counts comparable across different research fields and time. Different scaling factors have been proposed in the literature: the arithmetic mean, the geometric mean, the median, the z-score, etc. Due to the (right) skewness of citations' distribution it seems preferable to use the median as a scaling factor. Empirical evidence from bibliometric studies suggests, however, that the arithmetic average seems the most effective scaling factor when the average is based on the publications actually cited and thus excluding those not cited from the calculation of the arithmetic average (i.e. Abramo et al. 2012a, b).[18] Scaling of citation counts is carried out by multiplying the citations of each publication by the chosen scaling factor that characterizes the distribution of citations of articles from the same academic field and the same year.

12.3.3.2 Alternative DEA Models for Assessing Research Productivity

The main purpose of using DEA to assess research productivity is to obtain, through an optimization procedure based on linear programming, *a posteriori* weights to aggregate research inputs and outputs in order to derive a single metric, by means of an efficiency score or a composite indicator, reflecting relative achievement (De Witte and Rogge 2010).[19] The *a posteriori* weights may be variable (i.e., unit-specific) or common, and may or may not reflect (at least partially) experts' or stakeholders' opinions.

The flexible (unit-specific) weights resulting from conventional DEA models reflect its underlying assumption that each evaluated unit is allowed to choose, under certain regulatory conditions, its own set of input and output weights in order to show it in the best possible light relative to other units. It is thus able to exaggerate its own advantages and at the same time to downplay its own weaknesses in order to obtain the maximal possible evaluation score. But if after that it is

[18] Abramo et al. (2012a, b) also provided empirical evidence indicating that rankings of individual researchers obtained under different scaling factors (i.e., average, median, cited papers average, cited papers median) do not show significant discrepancies.

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

still weaker relative to other units in the sample this cannot be put down to the choice of input and output weights. On the other hand, some authors have argued that comparison and ranking is meaningful only if it is conducted on common grounds and thus favour the use of common but not necessarily equal weights across different outputs. Several variants/modifications, such as common weights DEA and average cross efficiency, have been used for such purposes (e.g. see Oral et al. 2014). Lastly, a combination of *a posteriori* and a priori (i.e., model and experts/stakeholders based) weights may also be possible. Two rather distinct approaches have been used in this respect: peer appraisal by means of cross efficiencies and value judgment DEA. In the former case, the value norms (i.e. DEA weights) of all evaluated units are taken into account when assessing the performance of each unit. In the latter case, the DEA weights assigned to (some or all) inputs and outputs are constrained to satisfy a priori restrictions in order to eliminate the possibility of assigning zero values to particular inputs and/or outputs and more generally to ensure DEA weights accord with intuition.

The second methodological aspect that has to be considered at the outset of the evaluation process is whether resources related to research activities will be taken into account or not. This refers to the choice between measuring efficiency or effectiveness. The former compares the outcome(s) of the research related activities relative to the resources employed for this purpose while the latter compares only the outcome(s) of the research related activities and not the means to achieve them. Conventional DEA models may be used to measure efficiency of research activities while measuring effectiveness is equivalent to the construction of composite performance indicators, which can be done using either DEA-based models such as the benefit-of-the-doubt, (BoD), (e.g. Cherchye et al., 2007) or linear programming models (e.g., Kao and Hung 2003).

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

12.3.3.3 Measuring Efficiency and Effectiveness of Research Activities

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

On the other hand, effectiveness of research activity can be estimated by means of two seemingly similar models that share a common feature: they account only for the output side and thus acting as output aggregator functions. In the input side they rely on Koopman's idea of a person who has at his/her disposal a unitary quantity of an aggregated input that is used for research activities. These two models are the BoD and the Kao and Hung (2003) (K&H) model, which gained increasing popularity in recent years as models used to construct composite indicators. Based on Karagiannis and Paschalidou (2014) we next present and contrast these two models under three different specifications of output weights: variable, common, and restricted.

The BoD model is essentially a tool for aggregating linearly quantitative performance sub-indicators into a single composite indicator when the exact weights are not known a priori (Cherchye et al. 2007). For each evaluated unit, it does so by implicitly assigning less (more) weight to those sub-indicators or performance aspects that the assessed unit is a relatively weak (strong) compared to all other units in the sample. Moreover, the estimated weights are allowed to vary across units and time.

In technical terms, the BoD is a benchmarking model that has a DEA-type structure in the sense that the composite indicator is defined by the ratio of actual to benchmark performance, both of which are given by the weighted sum of the sub-indicators considered. Since the composite indicator is designed to take values in the [0,1] interval, benchmark performance attains by construction the maximum value of one (Cherchye et al. 2007). In determining actual overall performance, the weights are selected in such a way as to maximize the value of the composite indicator of the evaluated unit. This in turn guarantees that any other weighting scheme would worsen the ranking of this unit. Moreover, when these weights are used by any other unit in the sample would not result in a composite indicator greater than one. The resulting weights are determined endogenously by solving for each evaluated unit problem (12.11).

$$\begin{aligned} I^k &= \max_{w_i^k} \sum_{i=1}^{N} w_i^k I_i^k \\ s.t. &\sum_{i=1}^{N} w_i^k I_i^j \leq 1, \qquad j = 1, \ldots, K \\ w_i^k &\geq 0, \qquad\qquad i = 1, \ldots, N \end{aligned} \tag{12.11}$$

where $\mathrm{I}_\mathrm{i}^\mathrm{k}$ is the ith sub-indicator of the kth unit, the higher the value the better, $\mathrm{w}_\mathrm{i}^\mathrm{k}$ are the weights to be estimated, *j* is used to index units and *i* to index sub-indicators which in our case correspond to different research outputs (i.e., types of publications, citations, patents).

The BoD model is equivalent to the multiplier form of the Charnes et al. (1978) input-oriented, constant returns to scale (CRS) DEA model when there is a single constant input that takes the value of one for all evaluated units.[20] Based on this, the dual formulation of the BoD model is given as (12.12).

$$
\begin{aligned}
& I^k = \min_{\lambda_j^k} \sum_{j=1}^{K} \lambda_j^k \\
& s.t. \sum_{j=1}^{K} \lambda_j^k I_i^j \geq I_i^k, \qquad i = 1, \ldots, N \\
& \lambda_j^k \geq 0, \qquad\qquad\quad j = 1, \ldots, K
\end{aligned}
\tag{12.12}
$$

where λ refers to intensity variables. This implies that the value of the composite indicator is in fact equal to the sum of the intensity variables. From the inequality constraints on the intensity variables it is clear that the BoD model exhibits constant returns to scale.

On the other hand, the Kao and Hung (2003) model has a similar structure in the sense of deriving a set of *a posteriori* weights that maximize the value of a composite research performance indicator but now under the assumption that this set of weights satisfies for each evaluated unit an adding-up/normalization constraint. The K&H model is written as (12.13):

$$
\begin{aligned}
& E^k = \max_{u_i^k} \sum_{i=1}^{N} u_i^k I_i^k \\
& s.t. \sum_{i=1}^{N} u_i^k = 1 \\
& u_i^k \geq 0, \qquad\qquad i = 1, \ldots, N
\end{aligned}
\tag{12.13}
$$

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

Besides these differences, Kao et al. (2008) have shown that the two models are related to each other as long as the set of sub-indicators to be aggregated are normalized at the outset to be within the [0,1] interval; that is, $0 \leq \mathrm{I}_\mathrm{i}^\mathrm{k} \leq 1 \ \ \forall\, \mathrm{i} = 1, \ldots, \mathrm{N}$. In this case one can verify that $\mathrm{E}^\mathrm{k} = \mathrm{I}^\mathrm{k}/\mathrm{W}^\mathrm{k}$ where $\mathrm{W}^\mathrm{k} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{w}_\mathrm{i}^\mathrm{k}$ and $\mathrm{u}_\mathrm{i}^\mathrm{k} = \mathrm{w}_\mathrm{i}^\mathrm{k}/\mathrm{W}$. This implies that the K&H model delivers values of the composite indicator that are close but

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

not always equal to those suggested by the BoD model. More importantly, Karagiannis and Paschalidou (2014) note that while from the BoD weights we can derive the weights implied by the K&H model the converse is not possible. This limitation of the K&H model is related to the type~~s~~ and number of constraints that it involves. On the other hand, for this same reason, the K&H model is computationally less demanding. Lastly, at present the K&H model has unknown aggregation properties and thus we cannot move the analysis of research productivity from the individual to institution (i.e., department or university) level in a theoretically consistent way.

In contrast, such an aggregation rule for the BoD model has been developed by Karagiannis (2016) within the framework of aggregate efficiency scores. In particular, it has been shown that the arithmetic average (in (12.14)) is the theoretically consistent aggregation rule for the BoD model.

$$I = \frac{1}{K}\sum\nolimits_{k} I^{k} \tag{12.14}$$

Thus, the aggregate composite performance indicator equals the simple (un-weighted) arithmetic average of the estimated individual composite indicators. This results from the single constant (unitary) input structure of the BoD model and the denominator rule (Färe and Karagiannis 2013) stating that consistency in aggregation of ratio-type performance measures, including efficiency indices, is ensured as long as the weights are defined in terms of the variable being in the denominator.[21] For an input-oriented model such as the BoD, these will be actual cost or input shares. But since all evaluated units have the same amount of (one unit) and face the same price for the single input, the share weights become equal to 1/K. In terms of research activity, this result implies that a department's research productivity can be simply estimated by means of the average research productivity of its faculty members.

Regarding now the estimation of research effectiveness in terms of common instead of variable weights, which according to Kao and Hung (2005) and Wang et al. (2011) among others have the advantage of making it possible to compare and rank the performance of all evaluated units and not only classify them as efficient or inefficient, both the BoD and the K&H models possess special features.[22] First, for the BoD model Karagiannis and Paleologou (2014) have shown that common weights are related to average cross efficiency, which is of particular interest in assessing research productivity (Oral et al. 2014) as it provides the basis of giving the right to every faculty member to have a "say" about the performance of other faculty members in the same institution. In particular, average cross efficiency in the BoD model is based on a set of common weights given by the simple arithmetic average of weights obtained from the self-appraisal version of the model, i.e., the

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

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

one discussed above. On the other hand, a common set of weights in the K&H model can be obtained by applying Kao and Hung (2005) compromise solution. That is by running a linear ordinary least squares regression (not including an intercept term) of the composite indicator obtained from the conventional form of the model, on the set of sub-indicators under the restriction that the estimated parameters sum up to one.

Finally, a set of weights on which a research productivity assessment may be carried out is that reflecting value judgment. For the BoD model, this is incorporated in terms of weights restrictions in the multiplier form of the model. Several types of weights restrictions have been used for this purpose including pie shares (e.g. Cherchye et al. 2007) and partial descending ordering (i.e., $\mathrm{w}_1^{\mathrm{k}} > \mathrm{w}_2^{\mathrm{k}} > \mathrm{w}_3^{\mathrm{k}} > \ldots$). The latter is a case of particular interest for the K&H model because then as Ng (2007, 2008) has shown there is no need to estimate the composite performance indicator by means of linear programming but rather to compute it based on partial averages; that is, the composite indicator is given as in (12.15), where i is used to index sub-indicators.

$$Max\left\{ I_1^k, \frac{\sum_{i=1}^{2} I_i^k}{2}, \frac{\sum_{i=1}^{3} I_i^k}{3}, \ldots \right\} \tag{12.15}$$

Lastly, the BoD model, as a DEA-type model, can also be used to examine research productivity over time by means of the corresponding technology-based (i.e. Malmquist or Hicks-Moorsteen) indices, an aspect of performance evaluation that cannot be done with the K&H model. For the BoD model, Karagiannis and Lovell (2016) have shown that *first,* the Malmquist and Hicks-Moorsteen productivity indices coincide, they are multiplicatively complete,[23] and the choice of orientation for the measurement of productivity change does not matter. *Second,* there is a unique decomposition of the sources of productivity change containing three independent components, namely technical efficiency change, neutral technical change and output biased technical change. *Third,* the aggregate output-oriented Malmquist productivity index is given by the geometric average between any two periods of the simple (un-weighted) arithmetic average of the individual contemporaneous and mixed period efficiencies.

12.3.3.4 An Illustrative Application

The empirical application is from Karagiannis (2016) who applied the BoD model to evaluate the research achievements of faculty members in the Department of Economics at the University of Macedonia, Greece during the period 2000–2006. In

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

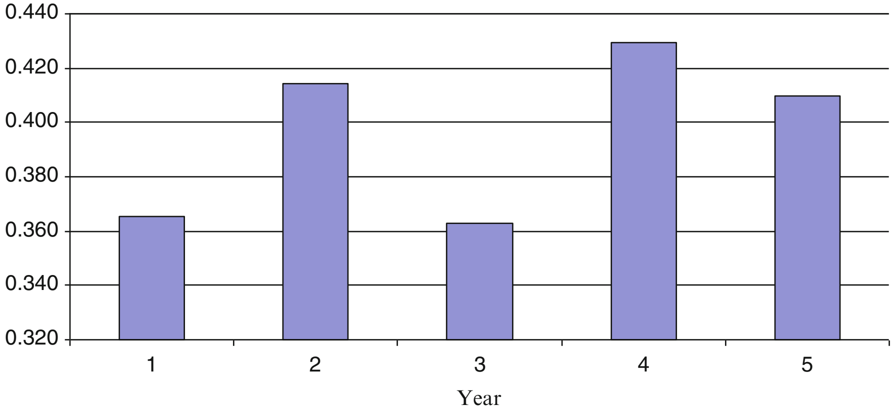

Fig. 12.9 Average annual technical efficiency: Source Karagiannis (2016)

the proposed setting the single constant input corresponds to each faculty member and two outputs were considered, namely, journal articles and all other publications, which are measured by whole numbers. As journal articles are considered all publications in outlets referenced in the Journal of Economic Literature and as other publications are considered papers published in journals not referenced in the Journal of Economic Literature, chapters in books and edited volumes. The relevant data reveal that, on average, each faculty member published almost one journal paper per year during the period 2000–2003 and there seems to be an improvement in research achievements as the annual average increases to somewhat above one during the period 2004–2006. The corresponding figures for other publications are well below one for the whole period, with a trend to decline significantly in the last 2 years. In addition, both kinds of publications are unevenly distributed between faculty members. There are a few faculty members with satisfactory achievements in journal article publications (more than two and a half on average per year) and one faculty member with similar performance in terms of other publications but most of them are around the departmental average. There were however two faculty members that had no journal article published during the whole period under consideration and with only one other publication each. Moreover, the achievements of newcomers are underestimated because of no entries in the data over the whole period under consideration. For convenience these two cases were disregarded and thus a total of 20 faculty members are included in the sample.

The average annual scores of technical efficiency (see Fig. 12.9), which reflect research efficiency at the department level, were found to be rather low and in the range of 0.36–0.43 indicating the relative heterogeneity in the achievements of faculty members.

12.3.4 Using DEA to Assess Administrative Services in Universities

The bulk of applications of DEA in Higher Education have focused either on assessments from a broadly academic perspective in terms of effectiveness of value added in teaching and research or in terms of economic efficiency at institution or department level. One DEA application which has addressed efficiency at university function level is that by Casu and Thanassoulis (2006) which has looked into university administrative services in the UK.

Universities as all publicly funded services are tasked to achieve value for money. Typically the focus when looking at universities has been on teaching and research. The allocation of resources between academic and non-academic departments has rarely, if ever, been subject to scrutiny. Yet, administrative expenditure is substantial. For example, in 1997/1998 expenditure on Administrative and Central Services (excluding services such as premises and catering) represented some 12 % of total UK higher education sector and nearly 30 % of expenditure on academic departments (Casu and Thanassoulis 2006). Using data for 1999/00, (Casu and Thanassoulis 2006) assessed expenditure on central administrative services (CAS) in UK universities, in order to identify the scope for potential savings in this area. A follow up study, as we will see later, used longitudinal data and the Malmquist index to measure the change in productivity in administrative services in UK Universities over the period 1999/00 to 2004/5.

University administrative services are organised at varying levels of devolvement to departments. The scope of functions (e.g. finance, personnel etc.) to be included within delineated units of assessment was decided with the aid of an advisory board of senior university academics and administrators (Casu et al. 2005). A follow up workshop was organised involving individuals at all levels of university administration, both academic and non-academic. A computer-mediated Group Support System (GSS) was used to home in on possible input-output variables to use in a DEA framework. The use of computer mediation was deemed beneficial by, for example, removing common communication barriers such as being interrupted, dominating discussants or a reluctance to share views. The software used was (‘JOURNEY’ – JOintly Understand, Reflect and Negotiate) to structure the discussion. Details of the facilitation leading to the delineation of the CAS unit of assessment and the related input-output variables can be found in (Casu et al. 2005). We summarise here the conclusions reached.

The services within a broad definition of CAS are illustrated in Fig. 12.10.

However, it was broadly agreed that in most universities services such as library, catering, residences etc. are self-contained both in organisational terms and in terms of data available. Hence they can be assessed separately. Thus the authors excluded self-contained functions to define CAS as the Core component in Fig. 12.10.

The main stakeholders were identified as:

- Students,
- Non-administrative Staff,

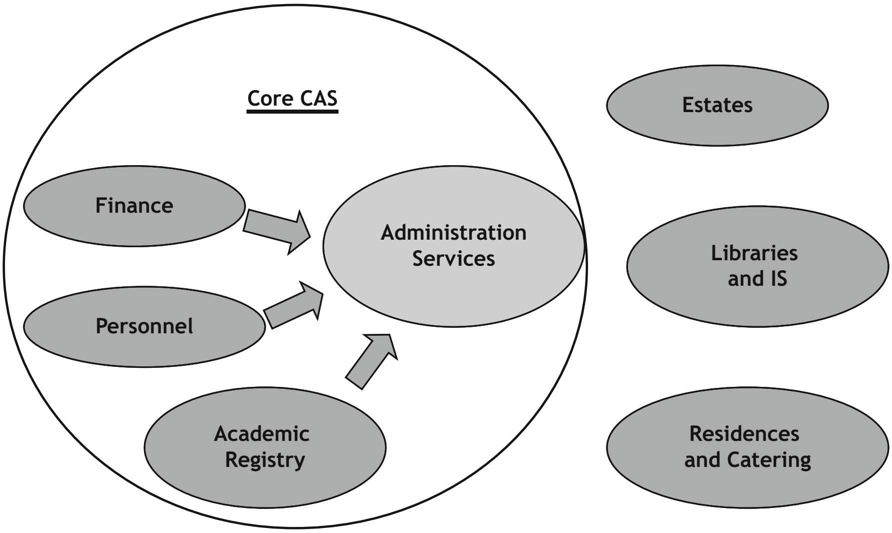

Fig. 12.10 Services included in CAS

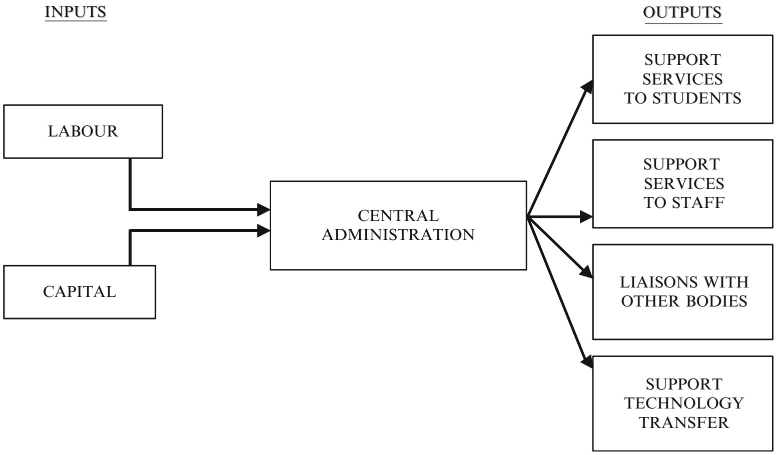
INPUTS
OUTPUTS
LABOUR
CAPITAL
CENTRAL ADMINISTRATION
SUPPORT SERVICES TO STUDENTS
SUPPORT SERVICES TO STAFF
LIAISONS WITH OTHER BODIES
SUPPORT TECHNOLOGY TRANSFER

Fig. 12.11 A conceptual framework of the CAS unit of assessment

- Suppliers,
- Funders,
- Community,
- Policy Makers and
- Educational and Industrial Partners.

This led to the conceptual model for identifying the input-output variables for CAS in Fig. 12.11 (Casu and Thanassoulis 2006).

Table 12.4 Inputs and outputs used in Casu and Thanassoulis (2006)

Inputs	Outputs
Total administrative costs (total administrative staff costs + other operating expenses)	Total income from students
	Total staff costs (minus total administrative staff costs)
	Technology transfer

Technology Transfer covers such activities as consultancy, joint ventures between the university and external organisations, setting up commercial companies to exploit the results of research and so on

Table 12.5 Modified input and output variables

Administrative cost measures (inputs)	Activity volume measures (outputs)
Administrative staff costs	Admin services for students Measure: *total income from students*
	Admin services for non-admin staff Measure: *non-administrative staff cost*
Operating expenditure on administration excluding staff costs	Services for research grants and other services rendered Measure: *income from these sources- using 3 year moving average*
	Services for research volume and quality Measure: *QR component of HEFCE income*

Secondary data as returned by Universities to the Higher Education Statistics Agency (HESA) for the year 1999/00 were used to operationalise the above framework for DEA purposes into the input-output variables in Table 12.4.

A total of 108 university administrations were assessed (England (86), Wales (7), Scotland (13) and Northern Ireland (2)). The assessment was run as an input minimisation, variable returns to scale DEA model. Considerable inefficiency was found – mean level of inefficiency 26.6 % – suggesting about a quarter of administrative expenditure in the universities assessed can be saved. This varied of course across institutions. Some 17 of them were identified as ‘efficient’ in the sense that no scope for savings could be identified in CAS services relative to other institutions.

The findings of the assessment need to be seen as indicative rather than definitive as the authors did not have detailed enough data as to what part of non academic staff costs in academic departments relates to the administration functions modelled. Moreover, the data returned to HESA by universities may not be fully comparable as institutions have some latitude in interpreting the data definitions.

The foregoing assessment was followed by an unpublished assessment of the change in productivity of administrative services of UK Universities between 1999/00 and 2004/5. (*The description here is drawn from the report submitted to the funders of the project*). Following the initial assessment using the 1999/00 data feedback from universities (a consultation event was attended by over 100 representatives from UK University administrations) the original input-output set was modified to the set shown in Table 12.5.

There are two main changes from the initial assessment; one is that two input variables are used, and the other is that a fourth output was added, (*QR component of research income*). One other change was that non administrative staff costs were adjusted to account for clinical staff present in some universities who have a substantially higher salary than non clinical staff.

The separation of Administrative Staff costs from Operating Expenditure (OPEX) was because experts could not agree whether the two types of resource can substitute for each other. The balance of view was mostly that they cannot substitute for each other, but data was not available to net out items of OPEX such as IT which can substitute for staff. There was a further pragmatic reason, apart from the issue of substitutability that led to the decision to focus on modelling the two inputs separately. Staff costs in comparison to OPEX are both more homogeneous across institutions and more clearly attributed to central and/or academic department administration. Modelling staff costs separately means that one can avoid contaminating this more clearly identifiable expenditure with the more heterogeneous and not easily attributable OPEX. This in turn made it possible to arrive at results which for staff expenditure would be much more reliable. In the event the authors ran three models, one using each input separately and a third using the two inputs jointly.

The fourth output added, *Quality-Related* (QR) research income, refers to the component of funding the institution receives for research as distinct from ad hoc research grants academic staff may secure through bidding. The QR component is based on the research quality rating achieved by each academic cost centre of that university in what at the time was the **research assessment exercise** (RAE). The amount also reflects the number of research active staff submitted by the University to the RAE. The QR output was intended to capture the administrative effort that academic research imposes on administrators over and above that already reflected in the other three outputs in the model. A weight restriction was used in the DEA model to ensure that one unit of QR funding was not deemed by the model to require more administrative (input) resource than an equal monetary unit relating to any one of the remaining three outputs. The QR component was also deflated to account for the fact that some disciplines receive higher funding than other disciplines for the same research quality rating depending on whether research requires laboratories etc.

Malmquist indices of productivity change were computed using the above input-output set with data for the period 1999/0-2004/5. The findings in summary were as follows:

- The scope for efficiency savings in administrative staff costs is of the order of 25 % in 1999/00 on average and it drops to about 20 % by 2004/5. There are some 20 units with good benchmark performance without any identified scope for efficiency savings. About half of these benchmark units are found to be consistently cost-efficient in all 6 years of our analysis. These units could prove examples of good operating practices in administration, which need to be identified by a careful study of the units beyond the data returned to HESA that were used.

- The picture in terms of OPEX efficiency is similar. The scope for efficiency savings is somewhat higher, between 20 % and 25 % on average each year. Again, some 20 units are benchmark and most of them are consistently so over time, offering the prospect of good operating practices.
- When taking administrative staff and OPEX as substitutable and using them jointly as inputs it is found that the scope for efficiency savings drops to about 15 % per annum in each one of the two inputs. The reduced scope found is because now benchmark units must offer low levels both on staff cost and OPEX rather than just on one of the two. As with the preceding two assessments here too it is found that a considerable number of units are benchmark in all 6 years which therefore could prove exemplars for other units to emulate. At the other extreme, some of the units with large scope for efficiency savings are so on all the 6 years suggesting persistent issues of expenditure control.
- Looking at productivity change between 1999/00 and 2004/5 a divergent picture is found between administrative single and two-input models. In the case of administrative staff taken as a self-contained resource it is found that there is on average a drop in productivity so that for given levels of the proxy output variables staff cost is about 95 % in 1999/00 compared to what it is in 2004/5, at constant prices. Looking deeper it is found that generally units do keep up with the benchmark units but it is the benchmark units that are performing less productively by almost 7 % in 2004/5 compared to 1999/00. The situation is not helped by a slight loss of productivity through scale sizes becoming less productive, by about 1.5 % in 2004/5 compared to 1999/00.
- In contrast when we look at OPEX as a self-contained resource or indeed at administrative staff and OPEX as joint resources it is found that productivity is more or less stable between 1999/00 and 2004/5. There is a slight loss of about 2 % but given the noise in the data this is not significant. What is significantly different between administrative staff on the one hand and OPEX or the two joint inputs on the other is that benchmark performance improves on average by about 8 % between 1999/00 and 2004/5. That is for given proxy output levels OPEX costs or joint OPEX and staff costs drop by about 8 % in 2004/5 compared to 1999/00. Unfortunately non benchmark units cannot quite keep up with this higher productivity. Also there is a slight deterioration again in scale size efficiency and the net effect is that despite the improved benchmark productivity, productivity on average is stable to slightly down between 1999/00 and 2004/5.
- As with cost efficient practices here too the analysis identified a small set of units which register significant productivity gains and others which register significant productivity losses between 1999/00 and 2004/5. An investigation of the causes in each case would be instrumental for offering advice to other units on how to gain in productivity over time and avoid losses. Such investigations need access to the units concerned beyond the HESA data used in the analysis.

12.3.5 Using DEA to Rank Universities

Universities are ranked in numerous ways, as a whole or by department by the popular press (De Witte and Hudrlikova 2013).[24] For example in the UK the Times, the Sunday Times and The Guardian are just some of the papers publishing so called League Tables. However, as universities are multi outcome entities the issue arises how to weight individual indicators of university performance such as research outcomes, teaching quality, employability of graduates etc. The DEA-based BoD model outlined earlier in Sect. 12.3.3 can be used to determine an endogenous weighting system, permitting each university to give weights that show it in the best light relative to other institutions. The rationale for this is that institutions should be allowed to have their own areas of excellence the ranking should be able to reflect this (e.g. see van Vught and Westerheijden 2010).

Let I_i^k be the index reflecting the position of university k on outcome i relative to other universities and let there be n outcomes on which universities are to be compared. It is assumed that the larger the value of I_i^k the better university k is relative to other universities. The BoD score I^k of university k can be computed through model (12.11) in Sect. 12.3.3. The BoD model chooses the weights w_i^k for each university k which maximize its ranking score I^k. The evaluated university 'can' choose the weights to maximize its 'aggregate' I^k. The weights are restricted not to permit any other university using those same weights to have a I^k value above 1. Because of this constraint the best performing universities will obtain a BoD score equal to 1. The rest of the universities will obtain a BoD score I^k lower than 1. The difference $(1 - I^k)$ expresses the shortfall in institutional attainments relative to other universities. The BoD scores can be used to rank universities.

De Witte and Hudrlikova (2013) deploy the foregoing method to rank the 200 universities originally ranked by The Times Higher Education Supplement (THES) in 2009. They use the variant outlined above which included repeated sampling with replacement and recomputation of ranks in the framework of the Cazals et al. (2002) approach. This mitigates the impact of outlying observations which might arise from measurement errors or from atypical observations (e.g., due to historic decisions). They use the same variables to rank universities as THES. These are:

- the reputation of the university (THES weight of 40 %)
- opinion on the university by employers (THES weight of 10 %)
- research excellence (THES weight of 20 %)
- staff-to-student ratio (THES weight of 20 %)
- international faculty (THES weight of 5 %)
- overseas students (THES weight of 5 %)

[24] This section is based on De Witte and Hudrlikova (2013).

They also control for contextual variables which might advantage or disadvantage some institutions. These are

- tuition fees,
- size of the university,
- research output (relative to size and faculty areas) and
- origin in an English speaking country.
- university autonomy.

The approach used to control for the foregoing contextual variables is that of De Witte and Kortelainen (2013). It was found that when universities are assessed accounting for exogenous background characteristics, European universities take the top 15 places. The original top universities according to the (unconditional) THES move from place 1 (Harvard University) to place 18, for Cambridge from 2 to 21 and Yale University from 3 to 24. These results suggest that it is important to control for contextual factors in comparing universities (see De Witte and Hudrlikova 2013 for an extensive discussion). Moreover, the results show that giving institutions the 'benefit of the doubt' on the multiple dimensions of institutions is important. As the nature of institutions differs, league tables should give credit to the relative strengths of each institution. DEA is a promising methodology for doing so.

12.3.6 Network DEA

All of the DEA models considered above are 'black box' in nature – that is, they consider the production process as one of conversion of inputs into outputs without further consideration of the mechanism by which this is done. An alternative approach, pioneered by Färe (1991) and Färe and Grosskopf (1996b), and more recently refined by Tone and Tsutsui (2009) involves constructing a network of nodes within the production unit; these nodes each have inputs and outputs, with some outputs of some nodes serving as inputs into other nodes. The overall efficiency of the production unit as a whole is then a function of the efficiency with which each node performs its role.

A network DEA of this type has been used to assess English higher education institutions by Johnes (2013). The network is illustrated in Fig. 12.12. The model shows two nodes. Node 1 uses measures of intake quality, student-staff ratio and per student spend to deliver degree results as an intermediate output. The degree results and the research reputation of the institution are inputs to Node 2 from which the ultimate outputs of the system are employability and student satisfaction. The results indicate that, while institutions are typically highly efficient in converting inputs into two of the outputs – student satisfaction and degree results – their performance in converting inputs into employability is less impressive, with only four institutions scoring above 90 % at the second node. This suggests that institutions looking to improve their overall efficiency might find quick wins in this area.

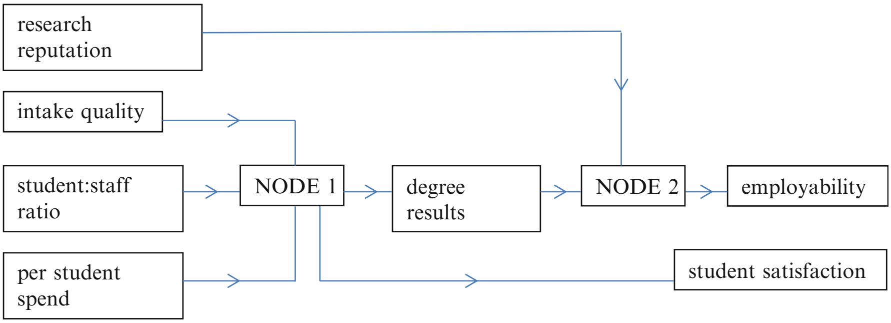

Fig. 12.12 A network DEA model

That said, the network structure used in this analysis is imposed by the analyst, and may be contentious. Different network designs are likely to lead to different results. Much work therefore remains to be done in the application of network models in the sphere of higher education.

12.4 Applications of DEA in Higher Education: Person Level

The foregoing DEA applications have focused on efficiency assessments broadly around the perspective of academic value added by a secondary school or a higher education institution. In this section we look at the use of DEA in assessing higher education academic staff on their key functions: research and teaching. De Witte et al. (2013a, b) discuss in a DEA model how the two activities relate, and whether there are economies of scope between teaching and research. This section focusses on teaching and research as two separate activities.

12.4.1 Assessing Academics on Research Output

Universities and colleges are increasingly interested in evaluating the performance of their academic staff, both in terms of teaching and research.[25] Current literature on research evaluation mainly employs single-criterion measures, such as reputational ratings gathered by peer reviews, number of publications in a predefined set of refereed journals, or citation counts (see De Witte and Rogge 2010). Recently, several authors have criticized such simplistic measures doubting whether they are able to accurately convey research performance. They argue the nature of research

[25] This section is based on De Witte and Rogge (2010).

is far too complex to be reflected in a single measure. A score to measure research should be multidimensional, control for exogenous characteristics and account for the different preferences by the different stakeholders and individuals. Therefore, 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:

1. How should one weight and aggregate the different research 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 research output criteria (i.e., equal/fixed weights)? Also, is it legitimate to apply a uniform set of weights to all evaluated researchers? Some researchers may focus on writing journal papers or books, while others in attracting research funding. Using the same weights for all researchers, could be seen as unfair within a research unit.
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, rank/tenure, time spent on teaching, department policy, etc.) may have a significant impact on the research performance of academic staff. Yet, traditional RES-scores do not account for differences in these uncontrollable conditions. Consequently, these scores are inherently biased towards researchers working under more favourable conditions.

The few studies which use multi-criteria instruments, calculate commonly a *global* RES-score for an individual as an arithmetic mean or a weighted sum of the researchers' performance on several criteria. They compute for researcher k the score as shown:

$$\begin{aligned} RES_k &= \sum\nolimits_{i=1}^{N} w_i^k I_i^k \\ \sum\nolimits_{i=1}^{N} w_i^k &= 1 \end{aligned} \tag{12.16}$$

where I_i^k is the number of outputs researcher k has in category i; w_i^k is the weight assigned to output category i for researcher k (with $0 \le w_i^k \le 1$ and $\sum\nolimits_{i=1}^{N} w_i^k = 1$); N is the number of output categories considered in the research evaluation. In studies where the RES-scores are computed as an arithmetic mean we have $w_i^k = 1/\mathrm{N}$. This implies that all aspects of research output are assumed to be of equal importance. In essence, an arithmetic mean RES-score corresponds to a measure where the publications are just counted over the different research output categories without any correction for their quality. When the RES-score is constructed as a weighted sum of publications with w_i^k varying over the different research output categories, this score corresponds essentially to a simple publication count with a correction for quality.

To account for the weighting issues in the construction of the research evaluation scores, De Witte and Rogge (2010) propose a specially tailored version of the '**benefit-of-the-doubt**' model. This data-driven weighting procedure has five important advantages compared to the traditional model as in Eq. (12.16).

First, for each evaluated researcher, weights for the various output criteria are chosen such that the most favourable 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 w_i^k which put the evaluated researcher k in the best possible light compared to his/her colleagues. As such, the research performance measure is relative. 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) cannot blame these poor evaluations to subjective or unfair weights. Any other weighting scheme than the one specified by the BoD model would worsen their RES-score relative to that of others. Second, the BoD model is flexible to incorporate stakeholder opinion (e.g., researchers, faculty administrators, 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, without pre-determining the exact weights. Third, 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. Fourth, 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, e.g. by using the robust order-*m* method of Cazals et al. (2002), adapting it to the 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.).

De Witte and Rogge (2010) applied the BoD approach on a dataset collected at the Department of 'Business Administration' of the Hogeschool Universiteit Brussel (Belgium) in the academic years 2006–2007 and 2007–2008. They argue their approach enables one to include different (potentially) influential conditions (outside the control of the researcher) into the built-up of the overall RES-scores. Hogeschool Universiteit Brussel (HUB) resembles in many ways 'new' (former 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 a research focus process initiated by the Government of the Flemish part of Belgium), Hogeschool Universiteit Brussel became increasingly research-oriented. The data set comprised research output data on all 81 research staff of the University. They added to this data on age, gender, doctoral degree, tenure, (official) teaching load, and (official) time for research. The data were further enriched with a questionnaire on the researcher's opinions and perceptions on research satisfaction and personal goals.

The core BoD idea is that output criteria on which the evaluated researcher performs well compared to his/her colleagues in the reference set Υ, 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) importance the evaluated researchers attaches to each criterion. For example, if, in comparison to his/her colleagues, the researcher under evaluation published a high number of papers in international journals this reveals that the researcher considers such publications to be of high importance. Consequently, his/her performances should weigh more heavily this criterion (i.e., high weight w_i^k). In other words, for each researcher separately, BoD looks for the weights that maximize (minimize) the impact of the criteria where the researcher performs relatively well (poorly) compared to other researchers. Hence, BoD-weights w_i^k in this sense are optimal and yield the maximal RES-score to the individual concerned (see (12.11) in Sect. 12.3.3).[26]

The BoD model for researcher k lets the data speak for themselves and endogenously selects those weights w_i^k which maximize his/her RES-score. Any other weighting scheme than the one specified by the BoD model would worsen the indicator $\mathrm{RES_k}$ for researcher k. This data-orientation is justifiable in the context of evaluating research performance 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 practice of using single-criterion measures or multiple-criteria as in (12.16) with or without a correction for the perceived quality. 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 a RES score based on pre-determined weights.

Note that the standard BoD model as in (12.11) grants evaluated researchers considerable leeway in the choice of their most favourable weights w_i^k. Only two constraints have to be satisfied. The first one is the ‘normalization’ constraint that ensures that all RES-scores computed with the evaluated researcher’s most favourable weights w_i^k, can at most be unity (or, equivalently, 100 %). Thus, we obtain $0 \leq RES_j \leq 1$ (j = 1,. . .,k,. . .,N) with higher values indicating better overall relative research performance. The second is the set which limits weights to be non-negative ($w_i^k \geq 0$). Apart from these restrictions weights can be chosen completely free to maximize the RES-score of the evaluated researcher vis-à-vis those of 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 no publications in any but one research output criterion, which may be generally not highly regarded, a researcher could place a high weight on that

[26] 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).

criterion and zero weights on all other criteria and achieve a high RES-score without violating the model restrictions. In such research evaluations, RES-scores reflect the researchers' performance on one single dimension. More generally it is possible for the BoD model to yield weights which deviate too much from what stakeholders (i.e., the faculty board, evaluated academics) would believe is appropriate. Without doubt, opponents of research evaluations will claim that RES-scores based on improper weights are not meaningful.

BoD models can be modified to incorporate weights restrictions as in traditional DEA models. Formally, this involves adding the general weight constraint (c) to the standard BoD-model:

$$w_i^k \in \mathrm{W_e} \quad \mathrm{i} = 1, \ldots, \mathrm{q} \text{ and } \mathrm{e} \in \mathrm{E} \tag{c}$$

with $\mathrm{W_e}$ denoting the set of permissible weight values defined based upon the opinion of selected stakeholders $e \in E$. It is crucial for the credibility and acceptance of RES-scores to define weight restrictions which reflect stakeholder opinions when available.

Using $w_i^k = \mathrm{w_{k,i}}$ De Witte and Rogge (2010) used the ordinal ranking of nine research output criteria, as agreed by stakeholders as in (12.17).

$$\mathrm{w_{k,1}} = \mathrm{w_{k,2}} \geq \mathrm{w_{k,3}} = \mathrm{w_{k,4}} \geq \mathrm{w_{k,5}} \geq \mathrm{w_{k,6}} = \mathrm{w_{k,7}} \geq \mathrm{w_{k,8}} = \mathrm{w_{k,9}} \geq 0.01 \tag{12.17}$$

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 I_i^k'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 I_i^k's.[27] Formally,

$$\begin{aligned} &(\mathrm{w_{k,1}} - \mathrm{w_{k,2}}) \times I_1^k \times I_2^k = 0 \\ &(\mathrm{w_{k,2}} - \mathrm{w_{k,3}}) \times I_2^k \times I_3^k \geq 0 \\ &\qquad\qquad \ldots \\ &(\mathrm{w_{k,6}} - \mathrm{w_{k,9}}) \times I_6^k \times I_9^k = 0 \\ &\mathrm{w_{k,i}} \times I_i^k \geq I_i^k \times 0.01 \qquad \forall\, \mathrm{i} = 1, \ldots, 9 \end{aligned} \tag{12.18}$$

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

In this adjusted version of the additional weight restrictions, a standard weight $w_{k,i} = 0$ for an output criterion i with $I_i^k = 0$ no longer forces other weights to be zero. In cases where one or both of the associated I_i^k's equal zero, the restriction becomes redundant and hence has no further influence on the other restrictions in (12.14). If none of the associated I_i^k's are zero, then the adjusted version of the weight restriction reduces to the original restriction as in (12.17).

De Witte and Rogge (2011) suggest two further improvements to the BoD model. First, they suggest to use a robust version such that the BoD model accounts for outlying observations without losing information due to removing such observations from the data set. Second, they suggest a conditional robust version. To do so, they apply the methodology suggested by De Witte and Kortelainen (2013), who extended the conditional efficiency framework to include discrete variables. The conditional efficiency framework compares like with like by computing for each observation a reference set with similar features. The BoD model is then estimated on this reference set with only comparable observations.

In a competitive context (e.g., for personnel selection decisions), by comparing researchers, the conditional RES-scores, which account for exogenous characteristics, can be deemed 'fairer' than unconditional RES-scores. Besides the employment conditions as retention, teaching load and research time the model used by the authors accounted for certain researcher background characteristics such as gender, age, PhD and being a guest researcher at University KU Leuven. Thus in effect their model controls for both exogenous factors (such as age, gender) and for factors which are exogenous to the researcher but not to the University (e.g. a university decision (e.g., hiring faculty without PhD, retention). The latter group of variables is interesting as it is at the discretion of the university. 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. Accounting for background variables, the conditional RES estimates increase dramatically. A larger group of researchers (75 %) becomes significant while the median researcher can improve her/his research performance by 21 % (see De Witte and Rogge 2010 for an extensive discussion).

12.4.2 Assessing Academics on Teaching

Students' evaluations of teaching (SETs hereafter) are increasingly used in higher education to evaluate teaching performance.[28] Yet, for all their use, SETs continue to be a controversial topic with teachers, practitioners, and researchers sharing the concern that SET scores tend to be 'unfair' as they fail to properly account for the impact of factors outside the teacher's control (De Witte and Rogge 2011). The reason for this concern is twofold. On the one hand, there are the numerous findings

[28] This section is based on De Witte and Rogge (2011).

in the academic literature which suggest that one or more background conditions (e.g., class size, subject matter, teacher gender, teacher experience, course grades, timing of the course) may have a significant influence on SET-scores (see, for instance, Feldman 1977; Marsh 1987, 2007; Marsh and Roche 1997; Centra and Gaubatz 2000; Marsh and Roche 2000). On the other hand, there is the practical experience from teachers themselves which indicates that some teaching environments are more conducive to high-quality teaching (and, hence, high SET-scores) while other environments make such a level of teaching more difficult.

De Witte and Rogge (2011) propose a DEA based Benefit of Doubt (BoD) approach, similar to that outlined above, for assessing academics on research, to assess them on teaching effectiveness. They construct SET-scores using a large array of single-dimensional performance indicators i (with $\mathrm{i} = 1,\ldots,\mathrm{N}$) where the weight placed on each indicator is derived through the BoD model. The conceptual starting point of BoD estimators, as we saw above, is that information on the appropriate weights can be retrieved from the observed data (i.e., letting the data speak).

The data consists of student assessments of the teacher on courses j ($\mathrm{j} = 1, \ldots, \mathrm{k}, \ldots, \mathrm{N}$) so that on course k the data on questionnaire item i, is I_i^k. For each teacher the BoD model assigns weights w_i^k to I_i^k so as to maximize the teacher's SET-score $\mathrm{SET_k}$. The model is as in (12.11) in Sect. 12.3.3.

To avoid problematic weight scenarios (zero or unrealistic weights), and to ensure the weights have intuitive appeal for teachers and students, additional weight restrictions are introduced in the basic model in the form of (12.17).

$$w_i^k \in \mathrm{W_e} \qquad \mathrm{i} = 1, \ldots, \mathrm{q} \text{ and } \mathrm{e} \in \mathrm{E} \tag{12.19}$$

where W denotes the set of permissible weight values based upon the opinion of selected stakeholders $e \in E$. For more details on this point see De Witte and Rogge (2011).

The authors further deploy the order-m method pioneered by Cazals et al. (2002) so as to estimate an *outlier-robust* SET score for each teacher. Moreover, they adapt the order-m scores so that they incorporate the exogenous environment (represented by R background characteristics $z_1, \ldots z_R$). This is done by drawing with replacement with a particular probability m observations from those observations for which $\mathrm{Z_{k,r}} \simeq \mathrm{Z}$. In particular, they create a reference group $\Upsilon^{m,z}$ from those observations which have the highest probability of being similar to the evaluated observation (similar in terms of the teaching environment in which the evaluated course was taught). The latter condition corresponds to conditioning on the exogenous characteristics $\mathrm{Z_{k,r}}$ (i.e., the teacher-related, student-related and course-related background characteristics). To do so, they smooth the exogenous characteristic Z by estimating a kernel function around $\mathrm{Z_{k,r}}$. Then they use the BoD model with the adapted reference set $\Upsilon^{m,z}$ to obtain estimates, labeled as $(\mathrm{SET_{k^m}}\left(I_i^k \middle| \mathrm{z}\right))$. These scores are not only robust to outlying observations (e.g., arising from measurement errors) but they also allow for heterogeneity arising from teacher, student and course characteristics.

De Witte and Rogge (2011) used the foregoing method to assess 69 different teachers of 112 college courses k ($k = 1,\ldots,112$) of the Faculty of Business Administration of HUB. Teachers who lecture several courses had several SET-scores, i.e. one for each evaluated course. Some 5513 students provided the feedback on the courses. The questionnaire comprised 16 statements to evaluate the multiple aspects of teacher performance. Students were asked to rate the lecturers on all items on a five-point Likert scale that corresponds to a coding rule ranging from 1 (I completely disagree) to 5 (I completely agree). The questions, covered 'Learning & Value', 'Examinations & Assignments', 'Lecture Organization', and 'Individual Lecturer Report'. For each course k ($k = 1,\ldots,112$) they calculated an average student rating I_i^k for each questionnaire item i ($i = 1,\ldots, 16$):

$$I_i^k = \frac{1}{S}\sum\nolimits_{S\in \ course\ k} I_{k,i,s} \tag{12.20}$$

where $I_{k,i,s}$ denotes the rating on question i of student s for the teacher who is lecturing course k. S is the number of students rating the course concerned. In terms of contextual variables $Z_{k,r}$, noted above, age of the teacher, gender, years of experience, whether or not he/she is a guest lecturer, whether or not the teacher received pedagogical training in the past and whether or not he/she has a doctoral degree were taken into account. Further, they included three background characteristics related to the students: the actual mean grade of the students in the class, the inequality of the distribution of the student grades (as measured by the Gini coefficient which can vary between 0 and 1, with a Gini coefficient of 0 indicating a perfectly uniform distribution and a Gini of 1 designating the exact opposite), and the response rate to the questionnaire. The latter captures the ratio of the number of people who completed the teacher evaluation questionnaire (i.e., S) to the (official) class size. Finally, two characteristics related to the course are included in the analysis: the class size and a dummy indicating whether the course is taught in the evening. They assessed teachers from three perspectives allowing progressively for the exogenous variables of teacher and student characteristics. The detailed models and findings can be found in De Witte and Rogge (2011).

12.5 Conclusion

This chapter has provided insights into the richness of DEA in education literature. By applying technical and allocative efficiency and productivity change techniques to educational data, policy relevant insights are obtained at both student level, school and system level. While this Chapter provided an overview of recent work, it is definitely not complete. De Witte and López-Torres (2015) and Johnes (2015) provide two complementary literature reviews on the efficiency in education literature. Their reviews show that many authors in various countries working with heterogeneous data sources are contributing to the literature. Despite these common efforts, there are still many aspects of efficiency in education to be explored.

De Witte and López-Torres (2015) argue that it is remarkable that the DEA (or Operations Research) literature studying education is still a distinct literature from the standard parametric 'economics of education literature'. The latter literature pays significant attention to the issue of causality, while this is not an issue in the DEA literature yet. Only few DEA studies acknowledge that the presence of endogeneity (e.g., due to omitted variable bias, measurement errors or selection bias) results in internal validity problems (notable exceptions are Ruggiero 2004; Haelermans and De Witte 2012; Cordero-Ferrera et al. 2013; Santín and Sicilia 2014). If the DEA literature on education aims to have more impact on the policy debate and on policy making, it should focus more on endogeneity and causal interpretations. The results from the DEA literature can now be easily criticised because of the lack of causal evidence. In relation to this, De Witte and López-Torres (2015) argue that the DEA literature should be more outward looking. Important developments in the economics of education literature, like experiments and quasi-experiments, have been largely ignored. There are few DEA studies that exploit experimental or quasi-experimental evidence (an interesting exception is Santín and Sicilia 2014). Yet, applying DEA to data from experiments or natural experiments in education might yield promising results. One may think of examining the efficiency of educational innovations, or changes at system level. Applying DEA to this type of data would help to bridge the gap between the DEA efficiency in education literature and the parametric efficiency in education literature.

References

Abramo G, D'Angelo CA (2014) How do you define and measure research productivity? Scientometrics 102(2):1129–1144

Abramo G, D'Angelo CA, Di Costa F (2010a) Citations versus journal impact factor as proxy of quality: could the latter ever be preferable? Scientometrics 84(3):821–833

Abramo G, D'Angelo CA, Solazzi M (2010b) National research assessment exercises: a measure of the distortion of performance rankings when labor input is treated as uniform. Scientometrics 84(3):605–619

Abramo G, Cicero T, D'Angelo CA (2012a) How important is the choice of scaling factor in standardizing citations? J Informetr 6(4):645–654

Abramo G, Cicero T, D'Angelo CA (2012b) A sensitivity analysis of researchers' productivity rankings to the time of citation observation. J Informetr 6(2):192–201

Abramo G, D'Angelo CA, Cicero T (2012c) What is the appropriate length of the publication period over which to assess research performance? Scientometrics 93(3):1005–1017

Abramo G, Cicero T, D'Angelo CA (2013a) Individual research performance: a proposal for comparing apples and oranges. J Informetr 7(2):528–539

Abramo G, D'Angelo CA, Rosati F (2013b) The importance of accounting for the number of co-authors and their order when assessing research performance at the individual level in the life sciences. J Informetr 7(1):198–208

Abramo G, D'Angelo CA, Di Costa F (2014) Variability of research performance across disciplines within universities in non-competitive higher education systems. Scientometrics 98 (2):777–795

Agasisti T, Johnes G (2009) Beyond frontiers: comparing the efficiency of higher education decision-making units across more than one country. Educ Econ 17(1):59–79

Agasisti T, Ieva F, Paganoni AM (2014) Heterogeneity, school effects and achievement gaps across Italian regions: further evidence from statistical modeling. MOX report number 07/2014. http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/07-2014.pdf. Dipartimento di Matematica "F. Brioschi", Politecnico di Milano, Via Bonardi 9 – 20133 Milano

Aigner D, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier production models. J Econometrics 6:21–37

Andersen P, Petersen NC (1993) A procedure for ranking efficient units in data envelopment analysis. Manag Sci 39(10):1261–1264

Arnold VL, Bardhan IR, Cooper WW, Kumbhakar SC (1996) New uses of DEA and statistical regressions for efficiency evaluation – with an illustrative application to public secondary schools in Texas. Ann Oper Res 66(4):255–277

Athanassopoulos A, Shale EA (1997) Assessing the comparative efficiency of higher education institutions in the UK by means of data envelopment analysis. Educ Econ 5(2):117–135

Ballou D, Sanders WL, Wright P (2004) Controlling for student background in value-added assessment of teachers. J Educ Behav Stat 29(1):37–65

Banker RD, Morey RC (1986) The use of categorical variables in data envelopment analysis. Manag Sci 32(12):1613–1627

Banker RD, Janakiraman S, Natarajan R (2004) Analysis of trends in technical and allocative efficiency: an application to Texas public school districts. Eur J Oper Res 154(2):477–491

Baumol WJ, Panzar JC, Willig RD (1982) Contestable markets and the theory of industry structure. Harcourt Brace Jovanovich, London

Bessent AM, Bessent EW (1980) Determining the comparative efficiency of schools through data envelopment analysis. Educ Adm Q 16(2):57–75

Bogetoft P, Nielsen K (2005) Internet based benchmarking. Group Decis Negot 14(3):195–215

Bradley S, Johnes G, Millington J (2001) The effect of competition on the efficiency of secondary schools in England. Eur J Oper Res 135(3):545–568

Burney NA, Johnes J, Al-Enezi M, Al-Musallam M (2013) The efficiency of public schools: the case of Kuwait. Educ Econ 21(4):360–379

Camanho AS, Dyson RG (2006) Data envelopment analysis and Malmquist indices for measuring group performance. J Prod Anal 26:35–49

Caporaletti LE, Dulá JH, Womer NK (1999) Performance evaluation based on multiple attributes with nonparametric frontiers. Omega 27(6):637–645

Casu B, Thanassoulis E (2006) Evaluating cost efficiency in central administrative services in UK universities. Omega 34(5):417–426

Casu B, Shaw D, Thanassoulis E (2005) Using a group support system to aid input-output identification in DEA. J Oper Res Soc 56(12):1363–1372

Cazals C, Florens J-P, Simar L (2002) Nonparametric frontier estimation: a robust approach. J Econometrics 106(1):1–25

Centra JA, Gaubatz NB (2000) Is there gender bias in student evaluations of teaching. J High Educ 71(1):17–33

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2(4):429–444

Charnes A, Cooper WW, Rhodes E (1981) Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through. Manag Sci 27 (6):668–697

Cherchye L, Kuosmanen T (2006) Benchmarking sustainable development: a synthetic meta-index approach. In: McGillivray M, Clarke M (eds) Perspectives on human development. United Nations University Press, Tokyo

Cherchye L, Moesen W, Rogge N, Van Puyenbroeck T (2007) An introduction to 'benefit of the doubt' composite indicators. Soc Indic Res 82(1):111–145

Cherchye L, De Witte K, Ooghe E, Nicaise I (2010) Efficiency and equity in private and public education: a nonparametric comparison. Eur J Oper Res 202(2):563–573

Chetty R, Friedman JN, Rockoff JE (2014) Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates. Am Econ Rev 104(9):2593–2632

Coleman Report (1966) The concept of equality of educational opportunity. Johns Hopkins University/US Department of Health, Education & Welfare, Office of Education, Baltimore/Washington, DC

Cordero-Ferrera JM, Crespo-Cebada E, Pedraja-Chaparro F, Santín-González D (2011) Exploring educational efficiency divergences across Spanish regions in PISA 2006. Rev Econ Apl 19 (3):117–145

Cordero-Ferrera JM, Santín D, Sicilia G (2013) Dealing with the endogeneity problem in data envelopment analysis. MPRA 4:74–75

Costa JM, Horta IM, Guimarães N, Nóvoa MH, Cunha JFe, Sousa R (2007) icBench: a benchmarking tool for Portuguese construction industry companies. Int J Hous Sci Appl 31 (1):33–41

Crespo-Cebada E, Pedraja-Chaparro F, Santín D (2014) Does school ownership matter? An unbiased efficiency comparison for regions of Spain. J Prod Anal 41(1):153–172

Daraio C, Simar L (2005) Introducing environmental variables in nonparametric frontier models: a probabilistic approach. J Prod Anal 24(1):93–121

Daraio C, Simar L (2007a) Advanced robust and nonparametric methods in efficiency analysis: methodology and applications. Springer, Dordrecht

Daraio C, Simar L (2007b) Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach. J Prod Anal 28(1):13–32

De Witte K, Hudrlikova L (2013) What about excellence in teaching? A benevolent ranking of universities. Scientometrics 96(1):337–364

De Witte K, Kortelainen M (2013) What explains the performance of students in a heterogeneous environment? Conditional efficiency estimation with continuous and discrete environmental variables. Appl Econ 45(17):2401–2412

De Witte, K. and López-Torres, L. (2015). Efficiency in Education. A review of literature and a way forward. Documents de treball d'economia de l'empresa – Working paper series Universitat Autonoma de Barcelona. 15/01, pp. 40. Journal of Operational Research Society. In press

De Witte K, Marques RC (2009) Capturing the environment, a metafrontier approach to the drinking water sector. Int Trans Oper Res 16(2):257–271

De Witte K, Marques RC (2010) Influential observations in frontier models, a robust non-oriented approach to the water sector. Ann Oper Res 181(1):377–392

De Witte K, Rogge N (2010) To publish or not to publish? On the aggregation and drivers of research performance. Scientometrics 85(3):657–680

De Witte K, Rogge N (2011) Accounting for exogenous influences in performance evaluations of teachers. Econ Educ Rev 30(4):641–653

De Witte K, Van Klaveren C (2014) How are teachers teaching? A nonparametric approach. Educ Econ 22(1):3–23

De Witte K, Thanassoulis E, Simpson G, Battisti G, Charlesworth-May A (2010) Assessing pupil and school performance by non-parametric and parametric techniques. J Oper Res Soc 61 (8):1224–1237

De Witte K, Rogge N, Cherchye L, Van Puyenbroeck T (2013a) Accounting for economies of scope in performance evaluations of university professors. J Oper Res Soc 64(11):1595–1606

De Witte K, Rogge N, Cherchye L, Van Puyenbroeck T (2013b) Economies of scope in research and teaching: a non-parametric investigation. Omega 41(2):305–314

Deutsch J, Dumas A, Silber J (2013) Estimating an educational production function for five countries of Latin America on the basis of the PISA data. Econ Educ Rev 36:245–262

Emrouznejad A, De Witte K (2010) COOPER-framework: a unified process for non-parametric projects. Eur J Oper Res 207(3):1573–1586

Färe R (1991) Measuring Farrell efficiency for a firm with intermediate inputs. Acad Econ Pap 19 (2):329–340

Färe R, Grosskopf S (1996a) Intertemporal production frontiers: with dynamic DEA. Kluwer Academic Publishers, Boston

Färe R, Grosskopf S (1996b) Productivity and intermediate products: a frontier approach. Econ Lett 50(1):65–70

Färe R, Karagiannis G (2013) The denominator rule for share-weighting aggregation. Mimeo

Färe R, Karagiannis G (2014) A postscript on aggregate Farrell efficiencies. Eur J Oper Res 233 (3):784–786

Färe R, Zelenyuk V (2003) On aggregate Farrell efficiencies. Eur J Oper Res 146(3):615–620

Färe R, Grosskopf S, Weber WL (1989) Measuring school district performance. Public Finance Q 17(4):409–428

Farrell M (1957) The measurement of productive efficiency. J R Stat Soc Ser A 120(3):253–281

Feldman KA (1977) Consistency and variability among college students in rating their teachers and courses: a review and analysis. Res High Educ 6(3):223–274

Flegg T, Allen D, Field K, Thurlow TW (2004) Measuring the efficiency of British universities: a multi-period data envelopment analysis. Educ Econ 12(3):231–249

Førsund FR (1993) Productivity growth in Norwegian ferries. In: Fried HO, Schmidt SS (eds) The measurement of productive efficiency: techniques and applications. Oxford University Press, New York

Fukuyama H, Weber WL (2002) Evaluating public school district performance via DEA gain functions. J Oper Res Soc 53(9):992–1003

Golany B, Storbeck JE (1999) A data envelopment analysis of the operational efficiency of bank branches. Interfaces 29(3):14–26

Goldstein H (1987) Multilevel models in educational and social research. Charles Griffin, London

Goldstein H, Huiqi P, Rath T, Hill N (2000) The use of value added information in judging school performance. Institute of Education, London

Gray J, Jesson D, Jones B (1986) The search for a fairer way of comparing schools' examination results. Res Pap Educ 1(2):91–122

Greene W (2005) Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. J Econometrics 126:269–303

Grosskopf S, Hayes KJ, Taylor LL, Weber WL (1997) Budget-constrained frontier measures of fiscal quality and efficiency in schooling. Rev Econ Stat 79:116–124

Grosskopf S, Hayes KJ, Taylor LL, Weber WL (1999) Anticipating the consequences of school reform: a new use of DEA. Manag Sci 45(4):608–620

Haegeland T (2006) School performance indicators in Norway: a background report for the OECD project on the development of value-added models in education systems. OECD, Paris

Haelermans C, De Witte K (2012) The role of innovations in secondary school performance – evidence from a conditional efficiency model. Eur J Oper Res 223(2):541–549

Hagen NT (2014) Counting and comparing publication output with and without equalizing and inflationary bias. J Informetr 8(2):310–317

Hanushek EA (1979) Conceptual and empirical issues in the estimation of educational production functions. J Hum Resour 14(3):351–388

Hanushek EA (1986) The economics of schooling: production and efficiency in public schools. J Econ Lit 24(3):1141–1177

Hanushek EA, Link S, Woessmann L (2013) Does school autonomy make sense everywhere? Panel estimates from PISA. J Dev Econ 104:212–232

Ibensoft Aps (2013) Interactive benchmarking: state-of-the-art in performance evaluation. From http://www.ibensoft.com/

Johnes G (1998) The costs of multi-product organisations and the heuristic evaluation of industrial structure. Socioecon Plann Sci 32(3):199–209

Johnes G (1999) The management of universities: Scottish Economic Society/Royal Bank of Scotland annual lecture. Scott J Polit Econ 46:502–522

Johnes J (2006a) Data envelopment analysis and its application to the measurement of efficiency in higher education. Econ Educ Rev 25(3):273–288

Johnes J (2006b) Measuring efficiency: a comparison of multilevel modelling and data envelopment analysis in the context of higher education. Bull Econ Res 58(2):75–104

Johnes J (2006c) Measuring teaching efficiency in higher education: an application of data envelopment analysis to economics graduates from UK universities 1993. Eur J Oper Res 174:443–456

Johnes J (2008) Efficiency and productivity change in the English higher education sector from 1996/97 to 2004/05. Manch Sch 76(6):653–674

Johnes G (2013) Efficiency in higher education institutions revisited: a network approach. Econ Bull 33(4):2698–2706

Johnes J (2014) Efficiency and mergers in English higher education 1996/97 to 2008/9: parametric and non-parametric estimation of the multi-input multi-output distance function. Manch Sch 82 (4):465–487

Johnes J (2015) Operational research in education. Eur J Oper Res 243(3):683–696

Johnes G, Johnes J (2009) Higher education institutions' costs and efficiency: taking the decomposition a further step. Econ Educ Rev 28(1):107–113

Johnes J, Johnes G (2013) Efficiency in the higher education sector: a technical exploration. Department for Business Innovation and Skills, London

Johnes G, Johnes J, Thanassoulis E, Lenton P, Emrouznejad A (2005) An exploratory analysis of the cost structure of higher education in England. Department for Education and Skills, London

Johnes G, Johnes J, Thanassoulis E (2008) An analysis of costs in institutions of higher education in England. Stud High Educ 33(5):527–549

Johnes J, Izzeldin M, Pappas V (2014) A comparison of performance of Islamic and conventional banks 2004-2009. J Econ Behav Organ 103:S93–S107

Johnson AL, McGinnis L (2011) Performance measurement in the warehousing industry. IIE Trans 43(3):220–230

Kao C, Hung H-T (2003) Ranking university libraries with *a posteriori* weights. Libri 53:282–289

Kao C, Hung H-T (2005) Data envelopment analysis with common weights: the compromise solution approach. J Oper Res Soc 56(10):1196–1203

Kao C, Hung H-T (2007) Management performance: an empirical study of the manufacturing companies in Taiwan. Omega 35(2):152–160

Kao C, Wu W-Y, Hsieh W-J, Wang T-Y, Lin C, Chen L-H (2008) Measuring the national competitiveness of Southeast Asian countries. Eur J Oper Res 187(2):613–628

Karagiannis G (2016) On Aggregate Composite Indicators, J Oper Res Soc (forthcoming)

Karagiannis G (2015) On structural and average technical efficiency. J Prod Anal 43(3):259–267

Karagiannis G, Lovell CAK (2016) Productivity measurement in radial Dea models with multiple constant inputs. Eur J Oper Res (fothcoming)

Karagiannis G, Paleologou SM (2014) Towards a composite public sector performance indicator. Asia-Pacific productivity conference, Brisbane, 8–11 Jul 2014

Karagiannis G, Paschalidou G (2014) Assessing effectiveness of research activity at the faculty and department level: a comparison of alternative models. Efficiency in education workshop, The Work Foundation, 19–20 Sept 2014

Kirjavainen T, Loikkanen HA (1998) Efficiency differences of Finnish senior secondary schools: an application of DEA and Tobit analysis. Econ Educ Rev 17(4):377–394

Kuosmanen T (2002) Modeling blank entries in data envelopment analysis. EconWPA working paper at WUSTL No. 0210001

Lazarsfeld PF, Henry NW (1968) Latent structure analysis. Houghton Mifflin, New York

Lin C-S, Huang M-H, Chen D-Z (2013) The influences of counting methods on university rankings based on paper count and citation count. J Informetr 7(3):611–621

Liu WB, Zhang DQ, Meng W, Li XX, Xu F (2011) A study of DEA models without explicit inputs. Omega 39(5):472–480

Lovell CAK, Pastor JT (1999) Radial DEA models without inputs or without outputs. Eur J Oper Res 118(1):46–51

Malmquist S (1953) Index numbers and indifference surfaces. Trab Estat 4:209–242

Mancebón M-J, Calero J, Choi Á, Ximénez-de-Embún DP (2012) The efficiency of public and publicly subsidized high schools in Spain: evidence from PISA-2006. J Oper Res Soc 63:1516–1533

Marsh HW (1987) Students' evaluations of university teaching: research findings, methodological issues, and directions for further research. Int J Educ Res 11:253–288

Marsh HW (2007) Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases and usefulness. In: Perry RP, SMart JC (eds) The scholarship of teaching and learning in higher education: an evidence-based perspective. Springer, Dordrecht, pp 319–383

Marsh HW, Roche L (1997) Making students' evaluations of teaching effectiveness effective: the critical issues of validity, bias, and utility. Am Psychol 52(11):1187–1197

Marsh HW, Roche L (2000) Effects of grading leniency and low workload on students' evaluations of teaching, popular myth, bias, validity, or innocent bystanders? J Educ Psychol 92 (1):202–228

Mayston DJ (2003) Measuring and managing educational performance. J Oper Res Soc 54 (7):679–691

Meyer RH (1997) Value-added indicators of school performance: a primer. Econ Educ Rev 16 (3):283–301

Mizala A, Romaguera P, Farren D (2002) The technical efficiency of schools in Chile. Appl Econ 34(12):1533–1552

Muñiz M (2002) Separating managerial inefficiency and external conditions in data envelopment analysis. Eur J Oper Res 143:625–643

Ng WL (2007) A simple classifier for multiple criteria ABC analysis. Eur J Oper Res 177 (1):344–353

Ng WL (2008) An efficient and simple model for multiple criteria supplier selection problem. Eur J Oper Res 186(3):1059–1067

O'Donnell C (2012) An aggregate quantity framework for measuring and decomposing productivity and profitability change. J Prod Anal 38(3):255–272

OECD (2008) Measuring improvements in learning outcomes: best practices to assess the value-added of schools. OECD Publishing, Paris

Oliveira MA, Santos C (2005) Assessing school efficiency in Portugal using FDH and bootstrapping. Appl Econ 37:957–968

Oral M, Oukil A, Malouin J-L, Kettani O (2014) The appreciative democratic voice of DEA: a case of faculty academic performance evaluation. Socioecon Plann Sci 48(1):20–28

Orea L, Kumbhakar SC (2004) Efficiency measurement using a latent class stochastic frontier model. Empir Econ 29(1):169–183

Pastor JT, Ruiz JL, Sirvent I (2002) A statistical test for nested radial DEA models. Oper Res 50 (4):728–735

Perelman S, Santín D (2011) Measuring educational efficiency at student level with parametric stochastic distance functions: an application to Spanish PISA results. Educ Econ 19(1):29–49

Portela MCAS, Camanho AS (2007) Performance assessment of Portuguese secondary schools. Working papers de Economia number 07/2007. https://ideas.repec.org/p/cap/wpaper/072007.html, Faculdade de Economia e Gestão, Universidade Católica Portuguesa (Porto)

Portela MCAS, Camanho AS (2010) Analysis of complementary methodologies for the estimation of school value added. J Oper Res Soc 61(7):1122–1132

Portela MCAS, Thanassoulis E (2001) Decomposing school and school-type efficiency. Eur J Oper Res 132(2):357–373

Portela MCAS, Camanho AS, Borges DN (2011) BESP – benchmarking of Portuguese secondary schools. Benchmarking 18(2):240–260

Portela MCAS, Camanho AS, Borges D (2012) Performance assessment of secondary schools: the snapshot of a country taken by DEA. J Oper Res Soc 63(8):1098–1115

Portela MCAS, Camanho AS, Keshvari A (2013) Assessing the evolution of school performance and value-added: trends over four years. J Prod Anal 39(1):1–14

Raudenbush S, Bryk AS (1986) Hierarchical models for studying school effects. Sociol Educ 59 (1):1–17

Ray A (2006) School value added measures in England: a paper for the OECD project on the development of value-added models in education systems. Department for Education and Skills, London

Ray A, Evans H, McCormack T (2009) The use of national value-added models for school improvement in English schools. Rev Educ 348:47–66

Ruggiero J (1998) Non-discretionary inputs in data envelopment analysis. Eur J Oper Res 111 (3):461–469

Ruggiero J (1999) Non-parametric analysis of educational costs. Eur J Oper Res 119:605–612

Ruggiero J (2004) Performance evaluation in education: modeling educational production. In: Cooper WW, Seiford LM, Zhu J (eds) Handbook on data envelopment analysis. Kluwer Academic Publishers, Boston

Sammons P, Nuttall D, Cuttance P (1993) Differential school effectiveness: results from a reanalysis of the Inner London Education Authority's junior school project data. Br Educ Res J 19(4):381–405

Sanders WL, Saxton AM, Horn SP (1997) The Tennessee value-added assessment system: a quantitative, outcomes-based approach to educational assessment. In: Millman J (ed) Grading teachers, grading schools: is student achievement a valid evaluation measure? Corwin Press, Inc. (Sage Publications), Thousand Oaks

Santín D, Sicilia G (2014) The teacher effect: an efficiency analysis from a natural experiment in Spanish primary schools. Efficiency in education workshop, The Work Foundation, 19–20 Sept 2014

Simpson G (2005) Programmatic efficiency comparisons between unequally sized groups of DMUs in DEA. J Oper Res Soc 56(12):1431–1438

Thanassoulis E (1996) A data envelopment analysis approach to clustering operating units for resource allocation purposes. Omega 24(4):463–476

Thanassoulis E (1999) Setting achievement targets for school children. Educ Econ 7(2):101–119

Thanassoulis E (2001) Introduction to the theory and application of data envelopment analysis: a foundation text with integrated software. Kluwer Academic Publishers, Boston

Thanassoulis E, Portela MCAS (2002) School outcomes: sharing the responsibility between pupil and school. Educ Econ 10(2):183

Thanassoulis E, Portela MCAS, Despić O (2008) DEA – the mathematical programming approach to efficiency analysis. In: Fried HO, Lovell CAK, Schmidt SS (eds) The measurement of productive efficiency and productivity growth. Oxford University Press, New York

Thanassoulis E, Kortelainen M, Johnes G, Johnes J (2011) Costs and efficiency of higher education institutions in England: a DEA analysis. J Oper Res Soc 62(7):1282–1297

Thieme C, Prior D, Tortosa-Ausina E (2013) A multilevel decomposition of school performance using robust nonparametric frontier techniques. Econ Educ Rev 32:104–121

Tone K, Tsutsui M (2009) Network DEA: a slacks-based measure approach. Eur J Oper Res Soc 197:243–252

Tsionas EG (2002) Stochastic frontier models with random coefficients. J Appl Econ 17:127–147

van Vught FA, Westerheijden DF (2010) Multidimensional ranking: a new transparency tool for higher education and research. Center for Higher Education Policy Studies (CHEPS), University of Twente, Enschede

Wang Y-M, Luo Y, Lan Y-X (2011) Common weights for fully ranking decision making units by regression analysis. Expert Syst Appl 38(8):9122–9128

Zhou P, Ang BW, Poh KL (2007) A mathematical programming approach to constructing composite indicators. Ecol Econ 62:291–297