Review of benchmarking methodologies

August 2018 | Produced by Alma Economics for the Office for Students

alma economics

Table of Contents

Acknowledgements

We are grateful to staff at the Office for Students and the Expert Advisory Group for their comments, in particular Mark Gittoes. We would also like to thank Alison Jones and Jennifer Summerton at the Higher Education Strategic Planners Association (HESPA) for their assistance administering the survey to higher education providers.

1. Context

1.1 Scope

This report is a review of benchmarking methodologies for the Office for Students (OfS), the regulator of higher education (HE) providers in England. It aims to contribute to the evidence base informing the HE sector about methods available for effective benchmarking[1]. The list of methods included aims to be as complete as possible. However, one method that is not assessed in detail is the OfS's existing benchmarking methodology. Since there is a parallel project examining the current method, we have not included an assessment of the existing approach in this review[2].

The assessment of methods in the report is generic in nature, in the sense that it does not consider existing HE data. However, certain data features that are prevalent in the English HE context are included in our assessment criteria such as how well methods deal with small numbers and atypical performance/extreme values. All methods included in the assessment could be applied to the HE context.

The aim of the report is to present a thorough assessment of the benefits and challenges associated with the application of each benchmarking methodology. The range of different methods available is large. With so much choice it is necessary for analysts to understand the pros and cons of different methods so they can produce the most suitable benchmarks for their needs – taking into account data availability, resources and how the benchmarks will be used.

The assessment takes the form of a *discussion* of the strengths and weaknesses rather than assigning scores and producing a definitive ranking. Producing a ranking of methods requires making judgements about the relative importance of different criteria which will vary for each organisation depending on their precise circumstances. The HE sector in England is diverse and each organisation uses benchmarking to suit its own requirements.

While every effort has been made to ensure our selection of methodologies and possible extensions has been thorough, in reality there are a vast number of permutations and combinations. For example, when carrying out data envelopment analysis (DEA), one can apply a number of extensions such as bootstrapping, a two-stage process, or dynamic application in a number of combinations. Rather than including every possible permutation, we have made judgements about the most useful methodologies to include in the assessment and the level of detail around their possible extensions to allow readers to get a sense of the options and their relative strengths and weaknesses.

A key part of this work has been an extensive literature review covering both benchmarking theory and applications, and we have included relevant references throughout the text providing further detail and insight on the methods covered.

The report is structured as follows:

Section 1 summarises the context and our approach to the review. Section 2 provides a review of the benchmarking methodologies, alongside a description of their applications. Our assessment strategy and the assessment of the different approaches are discussed in Section 3. Finally, Section 4 concludes.

1.2 What is benchmarking?

There are a large number of definitions of benchmarking available, though as noted by Meade (2007) this is largely due to differences in emphasis or application rather than a fundamental

[1] As recommended by the Higher Education Statistics Agency in 2010 (HESA 2010).

[2] Although the existing approach is part of the deterministic methods taxonomies discussed in Section 2.

disagreement about the concept of benchmarking.

An example of a definition is from Vlăsceanu *et al.* (2004) who note that benchmarking is a standardised method for collecting and reporting critical operational data in a way that enables relevant comparisons among the performance of different organisations or programmes. The purpose is usually to establish good practice, diagnose problems in performance, and identify areas of strength. From this evidence base, organisations can make informed decisions about improving working processes.

Considering the purpose and definition of a benchmark has implications for what methods constitute a 'benchmarking method'. The literature on benchmarking in a HE context is focussed on improving efficiency, and that is our focus in this report, though benchmarking can encompass a broader remit to be about any form of comparison. The European Centre for Strategic Management of Universities (ESMU, 2010) explains that the *'essence of a good benchmarking process is institutional learning'*. This encourages a broader concept of benchmarking, rather than simply improving the efficiency of HE providers. Therefore, while the majority of methodologies considered in the report deal with improving efficiency, there are some methods or applications of methods discussed which allow for a broader comparison across institutions.

While there is a generally agreed understanding of what benchmarking should aim to achieve (i.e. inform a decision-making process through comparison), there is no such thing as a 'natural' benchmark. The standard against which an organisation is compared will generally depend on circumstances. There will be times when comparing to a minimum standard is most appropriate, and others when comparing to an average of peers is more informative. Subjective judgement by the analyst designing the benchmark is always required. Therefore, assessing benchmarking methodologies is subjective and depends on the given context (see Section 1.3 and Appendix B).

Furthermore, while in the case of other types of quantitative models there are testable implications, any benchmark holds true by definition in reference to its underlying assumptions. For example, with forecasting models, we can compare predictions against outturns, but with benchmarking exercises there are no observable quantities against which our theoretical constructs can be evaluated.

While many aspects of benchmarking display significant variation, from the range of methods available, to the precise questions being answered by a benchmark, to how organisations are using benchmarks, a clear consensus lies in the fact that the purpose of benchmarks is to inform decisions. According to Meade (2007), once a benchmark has shed light on actions that lead to excellent performance, benchmarking should include *'the observation and exchange of information about those practices, the adaptation of those practices to meet the needs of one's own organisation, and their implementation'*. A theoretically elegant benchmarking methodology which does not help inform action may not be fit for purpose.

We hope this review supports the OfS and providers in the HE sector in their attempts to develop benchmarking approaches that support their strategic goals.

Box 1. Efficiency in the short and long run

A key feature of most benchmarking methods is that they attempt, for any given decision-making unit (DMU), to differentiate between 'inefficiency' (i.e. suboptimal performance due to factors deemed to be under the control of the DMU) and factors deemed to be difficult to change or completely outside the DMU's control.

For example, a benchmark of HE institutions may be adjusted to take into account the prior attainment of students, or the amount of funds available to carry out research – so that in assessing areas of 'inefficiency' these factors are deemed as being outside the institution's control. While this approach provides important insights in terms of where an institution may look to improve in the short run, over the long run DMUs can make strategic choices affecting most drivers behind their headline performance.

In other words, while comparing the research or teaching output of a more established university with a relative newcomer will yield limited insights into immediate 'inefficiencies' that need to be

addressed, it can also be misleading to ignore the fact the newcomer could adopt more fundamental changes over time (e.g. strategies to attract more funds, students with better prior qualifications, more experienced personnel, etc.) to impact the factors the benchmark is controlling for.

The above discussion highlights two key considerations to keep in mind:

- which question the approach is attempting to answer (e.g. what improvement may be possible in the short run, or how a DMU would perform if it had access to the resources of one of its peers, or which DMU is performing best in absolute terms, etc.)
- that none of the 'answers' provided by a benchmarking approach is the right one in an absolute sense, and that the aim is to provide a *useful* answer given the purpose of each particular benchmarking exercise.

1.3 Survey findings

Given the importance of context to benchmarking, we ran a survey of HE planning directors via the Higher Education Strategic Planners Association (HESPA). We note that this represents one, albeit important, group of HE benchmarking users and that surveying other groups (e.g. prospective students) may have led to different results.

The aims of the survey were to get an in-depth understanding about the contexts in which HE providers use benchmarks, as well as canvass opinions on how benchmarks could be improved. The key findings from the survey are discussed below, while more detailed results can be found in Appendix B.

The survey asked respondents how benchmarking helps inform decision-making and, separately, how it helps them achieve their strategic aims. The most common response to both questions was around using comparison to understand their own performance. This led to actions about targets and prioritisation. A selection of responses that illustrate the most common views are included below:

'Help us understand and interpret our performance.'

'…identify what we are doing well and where there is particular need for improvement.'

'Prioritisation of under-performing areas.'

'…gives an indication of how realistic a target to improve performance may be.'

There were a small number of institutions which did not feel benchmarks contributed to strategic aims because they did not consider the institution's specific circumstances.

Over 80% of providers felt that the HE benchmarks provided new information. When asked what aspects of the benchmarks users found helpful, the most frequent responses were that benchmarks are:

i) straightforward to understand (22%),

ii) ii) presented in a format that allows informed decision-making (19%),

iii) iii) contributed to the improved functioning of the HE provider (19%), and

iv) iv) transparent in how they are calculated (18%).

When asked about how current benchmarks could be improved, the most common responses were around improving the transparency of how they are calculated (41% of respondents), improved explanation around the results (23%) and making them more user-friendly in terms of allowing further analysis as well as clearer presentation (14%). Approximately 7% of respondents suggested improvements related to geographic context/weighting. Other improvements noted were around

data timeliness, better disaggregation, and the ability to add time trends.

2. Benchmarking methods

While it is common for practitioners to categorise benchmarking methodologies in different groups, no single categorisation appears to be uniformly followed in the literature. In order to provide structure to the review, we will group benchmarking methodologies into three wider families: deterministic, stochastic and frontier.

Broadly speaking, deterministic approaches generate performance measures that are assumed to be fully determined by the parameter values included in the corresponding model and the initial conditions. Stochastic models possess some inherent randomness, in the sense that the same set of parameter values and initial conditions will lead to an ensemble of different outputs (Pinsent *et al.*, 2016). Frontier methodologies directly establish the efficiency of each provider by comparing the distance between their observed output and the output which could be achieved if they were functioning at the most efficient level possible.

It is important to highlight that the current distinction between deterministic, stochastic and frontier methodologies is not watertight and it is primarily used to structure the discussion. For instance, even though stochastic frontier analysis (SFA) is technically a stochastic approach to efficiency analysis and DEA is a deterministic one, they are both also frontier methods and have been classified as such in the report. Given the importance of these two approaches in benchmarking applications, we feel it is best they are dealt with as a separate family.

2.1 Deterministic methodologies

2.1.1 Applied examples

Since the literature does not provide a formal taxonomy of deterministic approaches, to motivate and facilitate the creation of such a taxonomy we first review some practical examples. These consist of real world applications of deterministic methodologies used in the education sector as well as other parts of the public sector, such as healthcare and security.

Progress 8

In 2016, the Department for Education (DfE) introduced Progress 8, a new secondary school accountability system to measure the effectiveness of secondary schools in England (DfE, 2018). In the past, the main measure of school performance has been the percentage of pupils who achieved five or more A*– C grades in their General Certificate of Secondary Education (GCSE) exams, including maths and English[3].

The aim of Progress 8 at the individual level is to capture the progress a pupil makes from the end of primary school (i.e. Key Stage 2) to the end of the secondary school (i.e. Key Stage 4). It is a type of value-added measure[4], in the sense that each pupil's results are compared to the actual achievements of other pupils with similar prior attainment.

[3] It has been widely argued that as the old measure incentivised schools to focus on students at the C-D boundary, giving no reward to schools that improved their pupils from C grades. The same logic applied to those likely to gain grades E and below.

[4] According to Harvey (2004), the concept of 'value-added' relates to student achievement as growth in knowledge, skills, abilities, and other attributes that students have gained as a result of their experiences in an education system.

Estimation of Progress 8 at the school level requires obtaining the Attainment 8 and Progress 8 scores for every pupil individually.

Attainment 8 score

Each pupil's Attainment 8 score is the summation of the highest subject scores from four subject categories or baskets. The first basket is for English and the second is for maths. The grades achieved in these categories are double-weighted to reflect their importance. The third basket comprises three subjects that count in the English Baccalaureate (EBacc), including computer science, history, geography and languages. The final basket is for any other three GCSE subjects (including EBacc subjects used the previous basket) or any other non-GCSE subjects on the DfE approved list.

Pupils' Progress 8 and schools' Progress 8 scores

Progress 8 is calculated for individual pupils solely to derive each school's Progress 8 score. A pupil's Progress 8 score is defined as the difference between the actual Attainment 8 score and the average Attainment 8 score of all pupils nationally who had similar prior attainment, divided by ten. A pupil's academic starting point is also more commonly referred to as his/her prior attainment and from 2017 onwards, it is defined as the average of their Key Stage 2 results from primary school in English and maths.

A school's Progress 8 score is then calculated as the average of its pupils' Progress 8 scores. It gives an indication of whether, as a group, the pupils in this particular school made above or below average progress compared to pupils with similar prior attainment in other schools. Progress 8 scores generally fall somewhere between -1 and 1.

Notice that a negative Progress 8 score does not imply that pupils made no progress, but rather that pupils in the school made less progress than other pupils across England with similar academic starting points.

Confidence intervals

Progress 8 results are calculated for a school based on a specific cohort of pupils. To account for the uncertainty that comes from the fact that a particular school may have performed differently with a different set of pupils, 95% confidence intervals around the Progress 8 scores are provided. The confidence intervals act as a proxy for the range of scores within which each school's underlying performance measure can be confidently said to lie. The results of schools with a small cohort tend to have wider confidence intervals and this reflects the fact that the performance of a small number of pupils taking their Key Stage 4 exams can have a disproportionate effect on the school's overall results. Both the Progress 8 score and the confidence interval for a school should be considered when comparing with other schools, pupil groups or national averages.

Floor standard

The floor standard is the minimum standard for pupil attainment and/or progress that the Government expects schools to meet. From 2016, a school is considered to be below the floor standard if its Progress 8 score is below -0.5 and the upper band of the 95% confidence interval is below zero.

Even though the word benchmark is not explicitly mentioned, the floor standard constitutes a predetermined metric to which a school's performance is compared. In particular, it is the lowest performance threshold that each school is expected to achieve.

A comprehensive explanation of Progress 8 is outlined in *Secondary Accountability Measures* published by the Department for Education (2018).

University rankings

Over recent years, there has been an increasing tendency for media outlets in the UK and other countries to develop and publish rankings of universities and other HE institutions. In general, a ranking is a relationship between a set of items such that, for any two items:

- the first is either 'ranked higher than', 'ranked lower than' or 'ranked equal to' the second
- no pair of items is incomparable
- rankings generally have only an ordinal interpretation (i.e. only the order is important).

Examples of such rankings in the HE sector in the UK include the *Guardian University Guide*, *Times Higher Education Rankings*, *Complete University Guide* and *Times and Sunday Times Good University Guide*[5,6].University rankings are made with the use of quantitative performance indicators that cover a wide range of characteristics of those institutions, such as teaching and research activities, entry standards and student satisfaction. Even though the methodologies used to produce the rankings vary, they all have one element in common: they involve the calculation of overall performance scores by applying predefined weights to the relevant performance indicators.

To demonstrate the diversity of the methodologies mentioned above, Table 1 presents the measures on which rankings are built alongside their corresponding weights[7].

[5] The *Times and Sunday Times Good University Guide* requires a subscription to access further information.

[6] Examples of university rankings published outside the UK, include the *ARWU World University Rankings* (http://www.shanghairanking.com) and the *QS World University Rankings* (https://www.topuniversities.com).

[7] An exhaustive comparison or detailed presentation of these methodologies is beyond the scope of this text, but further insights can be obtained by visiting the methodology documents for each: *Guardian University Guide* (2016), *Times Higher Education Rankings* (2017) and *Complete University Guide* (2018).

Table 1. Performance indicators and weights

Guardian University Guide		Times Higher Education Rankings		Complete University Guide	
Indicator	**Weight**	**Indicator**	**Weight**	**Indicator**	**Weight**
NSS - Teaching	10%	Teaching reputation survey	15%	Entry standards	1
NSS - Assessment and Feedback	10%	Student-to-staff ratio	4.5%	Student satisfaction	1.5
NSS - Overall satisfaction	5%	Doctorate-to-bachelor's ratio	2.25%	Research quality	1
Continuation	10%	Doctorates-awarded-to-academic-staff ratio	6%	Research intensity	0.5
Value added	15%	Institutional income	2.25%	Graduate prospects	1
Student-staff ratio	15%	Research reputation survey	18%	Student-staff ratio	1
Expenditure per student	5%	Research income	6%	Academic services spend	0.5[8]
Entry scores	15%	Research productivity	6%	Facilities spend	0.5
Career prospects	15%	Research influence	30%	Good honours	1
		International-to-domestic-student ratio	2.5%	Degree completion	1
		International-to-domestic-staff ratio	2.5%		
		International collaboration	2.5%		
		Industry income	2.5%		

It is often argued that such rankings are not an appropriate tool to fairly demonstrate the relative performance of universities and other HE institutions (Goldstein and Spiegelhalter, 1996; Sarrico *et al.*, 1997; Turner, 2005). We discuss these claims in detail in the assessment section of this report.

Prison Rating System

The Prison Rating System (PRS), originally introduced in 2009 in the UK, was developed by the Criminal Justice Group (CJG) alongside the National Offender Management Service (NOMS). The framework was owned and managed by NOMS. However, NOMS has since been replaced by Her Majesty's Prison and Probation Service (HMPPS) who published the latest 2016/17 prison performance ratings (HMPPS, 2017a). These are the last set of prison ratings to be generated using the PRS, which has been replaced by the Custodial Performance Tool[9].

The PRS is a key performance indicator framework that calculates an overall band between 1 to 4 for every prison (HMPPS, 2017b). Each band represents the level of performance the prison is operating at:

1. Exceptional performance
2. Meeting the majority of targets
3. Overall performance is of concern
4. Overall performance is of serious concern.

[8] The *Guardian University Guide* (2016) uses a different set of weights for medicine, dentistry and veterinary sciences.

[9] At the time of writing no further information is available about this tool.

The framework has a hierarchical structure composed of four levels: i) overall PRS band, ii) domains, iii) drivers, and iv) measures. The performance of each measure feeds into a driver, the driver performance feeds into one of the framework's four domains (public protection, reducing reoffending, decency and resource management, and operational effectiveness) and finally each domain performance feeds into the overall PRS band.

The framework's domains are partially informed by the evaluations from Her Majesty's Inspectorate of Prisons (HMI Prisons) and the Measuring the Quality of Prison Life (MQPL) survey. The 2016/17 PRS band was based on the performance of 31 measures[10].

The relative importance of each of the framework's items (measure, driver and domain) in the overall PRS band is captured by predetermined weights. For example, the measure 'security audit' is considered the most important measure in PRS, with a relative weight of 9.5%.

Care Quality Commission rating system

In 2015, the Care Quality Commission (CQC), the independent regulator of health and adult social care in England, introduced a new system for rating care providers, based on the four-point grading scale used by Ofsted (Ofsted, 2015). According to this system, any trust will be given a rating of:

1. Outstanding
2. Good
3. Requires improvement
4. Inadequate.

To determine those ratings, CQC inspects the following eight core services in acute hospitals (CQC, 2018a): Urgent and emergency services; Medical care (including older people's care); Surgery; Critical care; Maternity; Services for children and young people; End of life care; and Outpatients.

However, when inspecting acute specialist trusts, CQC only selects the core services that are appropriate for the services the trust offers, while considering any additional services individually[11].

For each service reviewed, CQC asks key five questions (CQC, 2018b) about whether services are: Safe; Effective; Caring; Responsive; and Well-led.

Each service is given a rating on the scale mentioned above for each of the five key questions and CQC aggregates those ratings to produce an overall score for each service. After each service is rated, the trust as a whole is rated too. In order to obtain the overall score for each service and the trust, CQC applies a set of principles to ensure consistency in decision-making (CQC, 2018a). According to those principles, the five key questions should be weighted equally when aggregating as should the core services.

UK Performance Indicators

UK Performance Indicators (UKPIs) are a range of statistical indicators that provide information on the performance of HE providers in the UK, across areas of special interest, such as widening participation and graduate employment. HESA have published the UKPIs since 2002/03 on behalf of the OfS, the Higher Education Funding Council for Wales, the Scottish Funding Council and the Department of the Economy in Northern Ireland.[12]

The existing UKPIs approach uses benchmarks that are calculated from a weighted sector average

[10] For the complete list of measures, see HMPPS (2017c).

[11] For example, in trusts that specialise in treating children and young people, CQC also inspects two additional core services: neonatal and transition services.

[12] Between 1996/97 and 2002/03 UKPIs were published by the Higher Education Funding Council for England (HEFCE), which has recently been replaced by the Office for Students.

where weightings:

- are based on the characteristics of the students at the provider
- take into account some factors which contribute to differences between HE providers, such as entry qualifications, age, subject, mode of study and qualification aim.

However, it is out of the scope of this text to extensively cover the methodology currently applied. A parallel project by Professor David Draper is examining the current benchmarking methodology and any options for extension. A presentation of the key features of the existing approach can be found on the HESA website[13]. Draper and Gittoes (2004) offers a more extensive discussion of technical aspects of the methodology.

NACUBO benchmarking tool

The National Association of College and University Business Officers (NACUBO) is a membership organisation representing more than 1,900 colleges and universities across the USA. In 2007, NACUBO developed a simple online benchmarking tool that enables its members to compare results from three USA-wide surveys: the Tuition Discounting Study, the NACUBO-Common fund Study of Endowments and the Student Financial Services Survey[14] .The tool works by making comparisons against overall averages based on all participants as well as against averages of self-selected peer groups. Unfortunately, it is difficult to obtain further insights on NACUBO's benchmarking tool, since a subscription is required to gain to access to additional information.

Snowball Metrics

Snowball Metrics is an academia-industry initiative, originally started in the UK in 2010. It is owned by research-intensive universities in various countries, to ensure that its output is of practical use and does not rely on other metrics provider(s).

Snowball Metrics consist of a set of agreed and standard methodologies – or 'recipes' – for calculating research metrics (e.g. academic-corporate collaboration impact, research student to academic staff ratio, etc.) in a consistent way[15]. According to their creators, the methodologies are tested to ensure that they are supplier-agnostic i.e. the recipes do not depend on a particular data source or supplier but rather can produce metrics from any data source.

Snowball Metrics can be used to give information for a single provider but with agreement from other providers, institutions can compare their metrics to those of other providers. The Snowball Metrics Exchange service acts as a free 'broker service' for the exchange of Snowball Metrics between peer institutions who agree to share information with each other. Metrics exchange is completely voluntary, and it is entirely under each institution's control to accept or decline requests to share some or all Snowball Metrics with benchmarking clubs – groups of institutions which have agreed to exchange metrics with each other – or individual institutions.

Every institution is responsible for generating its own Snowball Metrics following the agreed methodologies, whether they are calculated using a bespoke system, in a spreadsheet, or in a commercial tool. It is noteworthy that the data underlying the metrics are not exchanged, only metric values are.

2.1.2 Deterministic methodologies taxonomy

The examples above enable us to distinguish three broad classes of deterministic benchmarking

[13] For more information, see the HESA website (https://www.hesa.ac.uk/data-and-analysis/performance-indicators/benchmarks).

[14] For more information, see the NACUBO website (https://www.nacubo.org/research/2018/nacubo-benchmarking-tool).

[15] For a detailed presentation of Snowball Metrics recipes, refer to Colledge (2017).

methodologies, depending on the number of benchmarks used.

2.1.2.1 Class 1 – Performance assessed using individual provider information only

The first class is composed of approaches where the performance of each provider is measured using information only from that provider. The resulting level of performance is defined in a way that allows for direct comparisons across providers (e.g. a 'good' provider). While no specific benchmark is calculated, the performance level of a particular provider can still be compared with the performance level of other providers. Consequently, this class still represents an approach to benchmarking.

Performance levels could be expressed by a metric, such as the overall performance scores used to produce HE institutions' rankings, or by performance bands constructed from combining weighted components, similar to the ones used in the PRS and CQC's framework (e.g. 'outstanding' providers). Snowball Metrics also fall into this class.

2.1.2.2 Class 2 – Individual provider compared to a single benchmark

The second class of these approaches is the case where providers compare themselves against a *single* benchmark that has been calculated. The single benchmark could be any mathematical construct against which to compare organisations, such as:

- the average, maximum, minimum and/or range of performance in the sector as a whole - where information from all providers is used
- the average, maximum, minimum and/or range of performance of a number of select providers where information from only a subset of providers is used (e.g. HE providers comparing themselves to the average of the Russell Group).

NACUBO's benchmarking tool constitutes an application of this approach.

2.1.2.3 Class 3 - each provider has their own benchmark

Finally, the third class includes methodologies for which *a single benchmark is generated for each provider* by using not only the provider's own data, but also data from other providers in the sector. The idea of creating individual benchmarks is common in frontier methods so we consider a deterministic approach equivalent. To emphasise the distinction between Class 2 and 3, in Class 2 we compare the provider to a benchmark, but the provider's own data is not a benchmark in itself. In this class, every provider has their own benchmark. Progress 8 and OfS's own methodology are examples of this class.

This final class of deterministic approaches can be further broken down if we consider the data taken into account when calculating the benchmarks. Here we identify two possible sub-categories: global vs local comparisons and direct or indirect standardisation. The choice in both instances will depend on the benchmarking user's aims.

When the relevant information captures characteristics of all providers, the resulting benchmarks are appropriate for global comparisons (i.e. a provider's performance is compared against the sector as a whole) while using data only from a selected group of peers (e.g. the Russell Group or the ex-94 group of universities or providers with similar characteristics) would yield metrics appropriate for more local comparisons.

Box 2. Direct and indirect standardisation

Direct standardisation involves calculating the observed overall progression rate of a provider when its progression rates across the potential confounding factors (PCFs) categories is

unchanged, but its distribution of students across those categories had instead matched the distribution of the students at the national level. This is equivalent to imagining that the mix of students in the provider was the same as that of all students nationally, not that of the provider's actual students.

Indirect standardisation involves calculating the observed overall progression rate of a provider when the distribution of students across the PCF categories is unchanged, but its progression rates across those categories were replaced by the national progression rates. This is like imagining how well the whole country would do with this provider's students.

Source: Draper and Gittoes (2004)

Further, as pointed out by Draper and Gittoes (2004), not accounting for the effects of PCFs – that is, variables that could be related to both an outcome and the factor that is considered to drive this outcome – could lead to potentially misleading results when calculating a provider's progression rate. In the context of HE, examples of PCFs include entry qualifications, age and gender. For example, it may not be correct to conclude that one university is more efficient at teaching than another by looking at degrees awarded, if it has more students with higher entry qualifications. We can divide the approaches in this class into those that use direct or indirect standardisation to adjust for these factors.

2.2 Stochastic methodologies

The range of stochastic, also known as statistical, models available is substantial and varies from simple to technically complex. Theoretically any of these models could be used to construct a benchmark. It would be unwieldy and unfocussed to review all statistical models, so this section concentrates on the stochastic approaches that are commonly used for benchmarking, with an emphasis on those that use regression residuals.

2.2.1 Ordinary least squares (OLS)

OLS is a classical technique for estimating the parameter(s) of a linear regression model that aims to capture the relation of a variable 'y' with explanatory variable(s) 'x'. For example, variable y could be the number of graduates of a particular university course and the explanatory variable x could be the teaching staff to student ratio. The OLS technique draws a line of 'best fit' to the observed data points by minimising the sum of the squares of the residuals (i.e. the vertical distance of the real data points to the fitted line). An illustration of the line of 'best fit' can be seen in *Figure 1*.

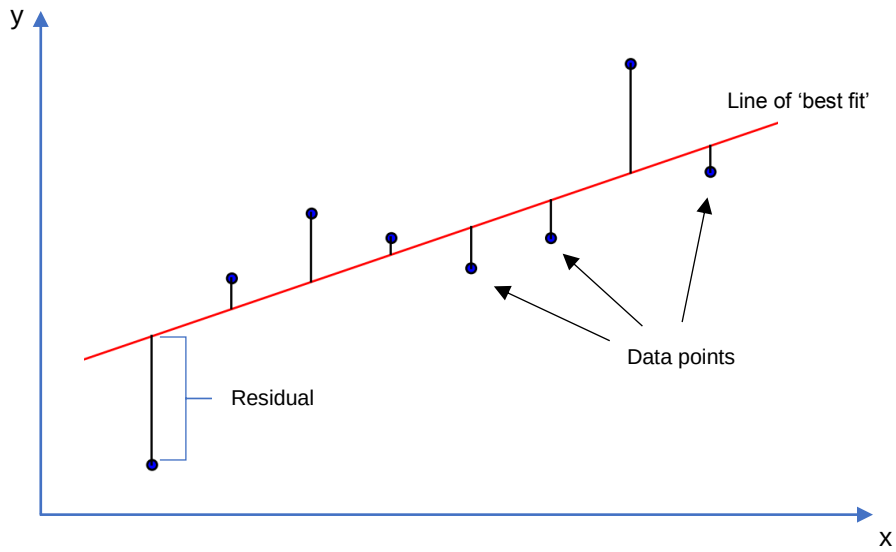


Figure 1. Real data points, line of 'best fit' and residuals

Given that the linear regression model is properly identified (i.e. all the relevant explanatory variables have been included), the regression residuals can be used to estimate whether providers are above or below the average efficiency level, and by how much. Average efficiency is defined as the estimated OLS regression line. Hence, a provider with a residual equal to zero is considered to be operating at the average estimated efficiency level. A provider with a positive residual (i.e. above the line), is operating above the average estimated efficiency in comparison to the other providers. In contrast, a provider with a negative residual is operating below the average estimated efficiency.

The application of OLS is not limited to the estimation of a linear regression model (i.e. where the line of best fit is a straight line). The method has also been extensively used with models specified to capture more complex relations. Usually, such models include multiple variables, which could be squared or raised to a higher power, variables in logs or even their lagged values.

Box 3. Application: OLS

Von Haldenwang and Ivanyna (2015) assess the performance of the tax system of 177 countries by examining the relation of each country's tax ratio to its level of development[16]. Based on the residuals of the OLS regression, the authors classify the countries as high and low tax performers. Surprisingly, their findings suggest that some low-income countries (e.g. Lesotho and Algeria) are among the top performers, while some high-income countries (e.g. Liechtenstein) fall into the low tax performing group.

2.2.2 Extensions to OLS

Corrected OLS (COLS)

Based on the idea that the best performer defines an 'efficiency frontier', the COLS model involves shifting the OLS line of 'best fit' to fit the best performing provider in the dataset. This adjustment allows all other providers to have a residual that is negative and can therefore be perceived as inefficient. Similar to the OLS model, efficiency estimates are calculated as the difference between each provider's real data point and the predicted values under COLS.

Modified OLS (MOLS)

OLS and COLS imply that every deviation from the frontier can only be caused by inefficiency. This assumption is likely not to be met in practice, as regression residuals also capture noise and measurement error. MOLS addresses this issue by shifting the OLS line of 'best fit' but to a lesser extent than COLS. The MOLS frontier lies between the OLS and COLS frontiers. Efficiency estimates under MOLS are calculated as the difference between each provider's real data point and the MOLS estimated efficiency frontier. The difference between the MOLS and the COLS can be interpreted as other noise/measurement error.

2.2.3 Value-added regression models

Value-added modelling can be defined as a category of statistical models that uses student achievement data to measure students' learning gain (Kim and Lalancette, 2013). This kind of modelling was initially considered as a method of measuring the teacher's influence on the students in a given school year. The idea of judging the effectiveness of teachers based on the learning gains of students was originally introduced in Hanushek (1971; 1992) and was further studied by Murnane (1975) and others.

In recent years, value-added modelling has been extended to enable researchers to examine the contribution not only of teachers, but also of schools and HE institutions. This has allowed for institutional ranking based on value-added scores. However, the models used in HE differ in many ways from the models used in secondary education, mainly because the data available in the two cases differ significantly (Kim and Lalancette, 2013).

The value-added models used in secondary education are developed based on longitudinal data that refer to the same students and subjects over time. Yet, longitudinal data are rarely found in the HE context. This is because students in HE are more difficult to track due to a relatively high level of mobility: they take leaves of absence, tend to change programmes or may even drop out of school. Due to this, most value-added models in HE use cross-sectional data and only few

[16] The logarithm of gross domestic product per capita has been used as a proxy for each country's development level.

longitudinal studies have been conducted[17].

In what follows, we present three cross-sectional approaches for value-added measurement used in HE. Even though the calculations used to produce these models vary, they all use test scores of students in two different year groups – the first year and final year – who take the test at the same time. All three models are based on regression residuals.

OLS linear regression-based approach

OLS linear regression models are used to capture whether students' average learning gain between first year and final year students in a given institution is near or above the expected test scores (i.e. what is observed at institutions admitting students with similar prior levels of attainment). In order to measure the expected test scores, these models involve regressing the current average test scores on the average scores on entry of first year and final year students respectively.

As the unit of analysis in these models is institutions rather than students, the dependent variable in each regression is the current average test score of every student in the institution. Each OLS linear regression equation generates its own residuals, which are the differences between the expected test scores as predicted by the regression model and the actual average test scores.

Moreover, since regressions for first year and final year students are conducted separately, residuals are obtained for both of those student categories. Then, the value-added score of an institution can be defined as the difference between the institution's first year students' residuals and the final year students' residuals.

Hierarchical linear models (HLM)

Similar to the OLS linear regression-based models presented above, HLM-based models also estimate the value-added scores of each institution looking at the difference between first year and final year students' residuals. HLM-based models, however, take into account that students are nested within institutions and that a student's academic growth could be affected by various institutional characteristics. The design adopts a two-level hierarchical approach (student level and institution level). These models also allow for full information at the student level to be used and therefore present a more accurate relationship between current test scores and entry test scores.

Specifically, the two levels of analysis are i) students (lower level) - with each individual student's test score being represented as a function of the student's academic ability scores on entry, ii) institutions (higher level). The relationship between student level variables is estimated based on institutional level information. The lower-level regression coefficients for each institution are the dependent variables, assumed to depend on institutional characteristics. In order to obtain first year and final year students' residuals, this multilevel model is estimated for every student category separately.

Box 4. Application: Value-added OLS and HLM approaches

Liu (2011b) compares the institutional value-added ranking that resulted from an HLM to the ranking produced by the OLS model in Liu (2011a). In order to obtain comparable results, the author used the same data for both models: test scores of 6,196 students from 23 HE institutions in the US, including 4,373 freshmen (first years) and 1,823 seniors (final year students).

The results suggest that institutional ranking was significantly different between these two methods. For example, some institutions went from being ranked towards the bottom to performing in the top 50%. This is a clear indication that estimates of institutional effectiveness

[17] See for example the Wabash National Study conducted in the USA (https://centerofinquiry.org/wabash-national-study-of-liberal-arts-education/).

can vary significantly depending on which method is used to calculate value-added scores.

HLM-based residual analysis approach

This approach also incorporates two levels of analysis to capture the institutional effects on the academic achievement of students. The main difference between the two methods presented above is that the HLM-based residual analysis approach – instead of using the difference in scores between first year and final year students – only compares the average test scores of final year students for each institution. The basic concept is that if the final year students at one institution achieve better test scores than is typical for institutions admitting students with similar entry qualifications, then greater learning gains have occurred.

In this case, the two levels of analysis are i) students (lower level) – where the final year student's current test scores are represented as a function of the entering student's academic ability score, ii) institutions (higher level). The relationship between the student-level variables is estimated based on both the average entry score for final year students and the current average test score of first year students[18].

Box 5. Application: *HLM-based residual analysis model*

Steedle (2012) works with data from two consecutive academic years – 2007/08 and 2008/09 – to compare the OLS with the HLM-based residual analysis value-added approach. The 2007/08 data included 12,898 freshmen (first year students) and 11,766 seniors (final year students) from 140 HE institutions, while the 2008/09 data included 24,252 freshmen and 14,578 seniors from 150 institutions. The author shows that even though the two models produce similar valued-added scores, the scores produced with the use of the HLM-based residual analysis approach are more reliable and stable across years.

Even though HLM-based residuals analysis increases the reliability of the institution effects compared to the OLS-based model, Steedle (2012) highlights that it is unlikely to be adequate *'for using value-added scores to make high-stakes decisions concerning educational efficacy, such as decisions about funding for public colleges and universities.'*

2.2.4 Other modelling frameworks

Instrumental variables

When the assumptions required for OLS to be unbiased or consistent are not satisfied, it can produce misleading results. In such cases, the use of instrumental variables, where available, enables the researcher to obtain consistent estimates for the model's parameters.

Box 6. Application: Instrumental variables

Schwartz and Zabel (2005) measure the efficiency of 62 public schools in New York by estimating the difference between the output (test scores) of each school and the output of other schools. In order to eliminate the potential endogeneity of school resources with the output, the authors use a set of lag variables as instruments, such as enrolment and the

[18] In the USA, both OLS and HLM-based residual analysis approaches have been widely used by the Council for Aid to Education (https://cae.org/) to produce the Collegiate Learning Assessment measures.

percentage of students with particular characteristics (e.g. eligibility for free lunches). Their results suggest that the best schools have lower teacher to pupil ratios as well as lower non-teacher expenditures.

Logistic model

Logistic analysis is a type of regression analysis used when the dependent variable is binary (e.g. win/lose, pass/fail or survived/died). The two possible dependent variable values are usually labelled as '0' and '1'. The goal of logistic regression is to find the values of the parameters that best describe the relationship between the binary variable and a set of independent variables. Rather than choosing parameters by minimising the sum of the squares of the residuals – as in OLS – this approach chooses the parameters that maximise the likelihood of observing the sample values.

In the context of benchmarking, logistic analysis is commonly conducted when estimating the performance of hospitals and other health care providers.

Box 7. Application: Logistic model

The Summary Hospital-level Mortality Indicator (SHMI) – published by the Health and Social Care Information Centre (NHS Digital) since 2011 – reports on mortality at the trust level across the NHS. The SHMI is defined as the ratio of the actual number of patients who die following hospitalisation at a trust to the number that would be expected to die on the basis of average England figures.

The expected number of deaths is estimated with the use of a logistic regression model that adjusts for the characteristics of the treated patients, including sex, age, current and underlying medical condition(s), year of discharge and method of admission[19].

In order to help stakeholders understand the SHMI, NHS Digital categorises the trusts into bands indicating whether a trust's SHMI is 'lower than expected', 'as expected' or 'higher than expected'[20].

Panel data

Panel data, sometimes called longitudinal data or cross-sectional time-series data, is a dataset in which the behaviour of individual units is observed over several time periods. These units can be countries, organisations (e.g. education providers), households or individuals (e.g. students) etc. Building such a dataset involves surveying a number of individual units at numerous points over time. Cross-sectional data can be seen as special cases of panel data, in the sense that they include information for many individual units, but only for one point in time.

Panel data analysis enables researchers to control for variables that are constant over time but cannot be observed or measured – such as cultural factors – or variables that change over time but not across all entities. Panel data is also suitable for multilevel or hierarchical modelling.

Designing panel surveys as well as conducting data collection across time is expected to be more resource intensive. Also, panel data can lead to some specific technical and practical issues (e.g. missing values) that are typically not present when working with cross-sectional data[21].

[19] Note that this method corresponds to indirect standardisation.

[20] For the latest version of the SHMI, as well as an exposition of the methodology according to which the indicator is calculated, the interested reader can refer to https://digital.nhs.uk/data-and-information/publications/clinical-indicators/shmi/current

[21] For a more extensive discussion, see Baltagi (2005).

Box 8. Application: Panel data

List and McHone (2000) use pollution data from 1986 to1997 to rank U.S. states according to environmental outputs across both air and water. The authors find that marginal performers using input-based rankings (e.g. Duerksen, 1983) – such as Wyoming and South Carolina – are among the top performers of the panel data ranking system. It is also shown that controlling for the level of state income causes significant change to the ranking position of some states.

2.4 Frontier methodologies

Frontier methodologies establish what 'best practice' looks like in terms of efficiency and plot where organisations lie in relation to the best practice frontier. A DMU's efficiency is a measure of how close that unit's observed output is to the output which could be achieved if it were producing at the most efficient level possible.

These methodologies build on the assumption that the process according to which an organisation transforms inputs into outputs can be captured by a production function. A production function is a mathematical formula that links the inputs (often factors of production, such as labour and capital) to the amount of output that can be produced. In the context of education, the production function represents the maximum output that can be achieved given the available resources. In the case of frontier methodologies in particular, this level of maximum output serves as a reference to obtain the relative efficiency of the organisations who fail to achieve it.

Box 9. Technical extension: The production function

An illustration of a production function is found in *Figure 2*:

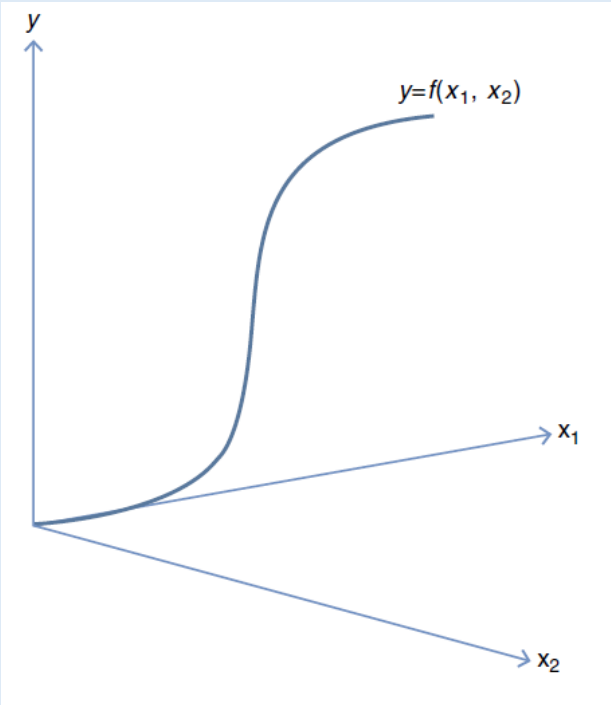


Figure 2. A production function with two inputs (x_1 *and* x_2) and one output (y)

Source: Titus and Eagan (2016)

If we consider that the organisation under examination is a factory that combines labour (x_1) and

specialised machinery (x_2) to produce smartphones (y), then the relation between the number workers and machines (inputs) and production size (output) is captured by the factory's production function.

The class of frontier models can be split by whether they are parametric or non-parametric models. The key difference between the two is the strength of assumptions about the distribution of data required in order to use the model. For example, parametric models require making an assumption about the distribution of the underlying data which are assumed to be defined by a set of parameters (e.g. because they follow a normal distribution they can be defined by the mean and variance). Non-parametric models assume the data cannot be defined by a fixed set of parameters but instead they use data to estimate the relationship. Non-parametric methods allow the relationship between a set of variables to take any functional form making this class of models more flexible.

Frontier models, using both parametric and non-parametric approaches, have attracted significant attention from researchers. For further detailed information about using frontier methods in the education sector (including primary and secondary phases) see Worthington (2001), Johnes and Johnes (2004), De Witte and López-Torres (2015), Aleskerov *et al.* (2017) and Johnes *et al.* (2017).

2.4.1 Non-parametric frontier methodologies

Non-parametric approaches have the benefit of not requiring a particular functional shape for the frontier, but they do not provide a general relationship relating output and input (i.e. there is no specified equation). Probably the most widely used non-parametric frontier method is DEA.

2.4.1.1 Data envelopment analysis

DEA is an approach for evaluating the performance of a set of organisations where the presence of multiple inputs and outputs makes comparisons difficult. The method identifies the best performing units in a comparable set and then determines the relative efficiency of each unit by assessing a DMU against only those that perform best. When the top performers are better than the DMUs under examination – either by making more output with the same input or making the same output with less input – then the DMU is inefficient. DEA makes it possible to identify where DMUs could improve.

DEA is a linear programming technique used to estimate the relative performance of homogeneous[22] organisational units/peer entities (DMUs) that use the same inputs to produce the same outputs. The method chooses weights (for both outputs and inputs) for each DMU, so that the weighed ratio of outputs to inputs is as large as possible given some constraints. It was initially developed by Charnes *et al. (*1978; 1979) while the background for the method is provided by Farrell (1957). DEA takes the observed input and output values to form a production possibility space, against which the individual units are compared to determine their efficiencies[23]. Note that those DMUs indicated as efficient are only efficient in comparison to the other DMUs in the set.

Input and output orientations

A wide range of DEA models have been developed to measure efficiency and largely fall into the categories of being either output-oriented or input-oriented:

Output-oriented models are used to examine whether the DMU under evaluation can increase its outputs while keeping the inputs at their current levels.

[22] The units are homogeneous in the sense that they that perform the same type of functions and have identical goals and objectives.

[23] DEA constructs the frontier on the basis of three assumptions: positive weights, non-discrimination of the units that are not dominated by any other unit, and the feasibility of linear combinations of the best performers (Storrie and Bjurek, 2000).

Input-oriented models are used to examine whether the DMU under evaluation can reduce its inputs while keeping the outputs at their current levels.

Coelli and Perelman (1999) show that the choice of orientation will have minor influences upon the scores obtained. However, Joumady and Ris (2005) point out that, when it comes to education, an output orientation may be more appropriate to the extent providers can be seen to be maximising output given a fixed set of resources (e.g. students' prior attainment)[24].

Constant and variable returns to scale and scale efficiency

The term 'returns to scale' refers to the production process according to which the units under examination transform inputs into outputs, and in particular in the rate of increase in inputs relative to an increase in inputs. When an increase in inputs results in the same proportional increase in outputs, this process exhibits constant returns to scale (CRS). However, if an increase in inputs does not cause the same proportional change in the outputs, the production process has variable returns to scale (VRS).

The Charnes *et al.* (1978) DEA model presented above performs well under the CRS condition (i.e. under the assumption that an increase in inputs results in the same proportional increase in output). If this condition does not hold, an additional constraint enables the DEA model to accommodate for the fact that production technology may exhibit VRS [25,26].

Scale efficiency or the scale effect refers to the extent in which a DMU is able to change its size toward the optimal size – defined as the area in which the CRS condition holds – in order to take advantage of returns to scale. This allows researchers to split estimates of efficiency into scale effects or other technical factors.[27]

Box 10. Application: Scale efficiency using DEA

Abbott and Doucouliagos (2003) obtain estimates of technical and scale efficiency for 36 Australian government universities by applying DEA and using various measures of inputs and output. Regardless of the output-input mix, their results suggest that Australian universities are operating at a fairly high level of efficiency relative to each other, although there is room for improvement in some of them.

Additive or slack-based DEA models

The output-oriented DEA models consider the possible (proportional) output augmentations while keeping the current levels of inputs. Similarly, the input-oriented DEA models consider the possible (proportional) input reductions while maintaining the current levels of outputs. Charnes *et al.* (1985) developed an additive DEA model which allows for possible input decreases as well as output increases simultaneously. The additive model is based upon input and output slacks. Input slack exists when it is possible to reduce an input without changing the level of output. Output slack exists when output can be increased without increasing the inputs. Another measure of efficiency is whether there is no input or output slack[28].

Additional constraints

[24] For a technical exposition of DEA as well as input-oriented and output-oriented DEA models, see Box A1 in Appendix A.

[25] It is noteworthy that Podinovski (2004) developed a model with selective proportionality where some indicators satisfy the CRS condition and the rest the VRS condition.

[26] For a more technical exposition, see Box A2 in Appendix A.

[27] See Box A3 in Appendix A for a technical exposition of scale efficiency in DEA.

[28] For a more technical exposition, see Box A4 in Appendix A.

The result of a DEA model is always a best-case scenario, in the sense that the approach provides a means by which weights may be self-determined and allows each unit to be assessed in its most favourable light. However, a potential problem with the best-case scenario is that DMUs manipulate the weights, including setting some weights to zero, to appear the most efficient but in practice the weights are not realistic (Allen *et al.*, 1997; Halme *et al.*, 1999). For instance, in the context of HE, it is often pointed out that weights should:

- be higher than a certain bound – to highlight the importance of particular factors and/or ensure that no factor is assigned a weight of zero
- be lower that a certain bound – to ensure that no weight is unrealistically high
- capture the relative importance between various inputs or outputs (e.g. teaching should have a higher weight than administrative services).

Box 11. Technical extension: Additional constraints

It is sometimes desirable to impose further restrictions in a DEA model. These additional constraints could be of several forms, including:

upper and lower bounds: $v_1 > 0.1, v_1 + v_2 < 0.4\ , v_1 + v_2 > v_3$

rankings: $v_1 > v_2\ or\ v_3 > \ v_1 > v_2$

Note that the additional constraints in the linear DEA program may transform to additional variables in the dual linear program and the calculation of the slacks could be affected as well[29].

2.4.1.2 Composite indicators with DEA

DEA can be used as a means of compiling various partial indicators into a single index, usually referred to as a synthetic or composite indicator. Composite indicators have increasingly been accepted as a useful tool for benchmarking, performance comparisons, policy analysis and public communication in many different fields. A number of applications in the construction of such indicators using DEA can be found in the literature: Hashimoto and Ishikawa (1993), Zhu (2001) and Murias *et al.* (2006) use composite DEA indicators to analyse economic welfare while Despotis (2005) recalculates the Human Development Index using DEA [30,31].

The following three features make the application of DEA attractive for constructing composite indicators:

1. Best performance in DEA is not a theoretical or abstract concept but is rather based on actual data
2. DEA respects the individual characteristics of the units and their own particular value systems, because it allows each unit to choose its weights individually
3. The flexibility in how weights are assigned to partial indicators may be graded or limited by introducing additional restrictions to the values of the weights. These limitations can be easily incorporated into the DEA model (see previous section).

Box 12. Application: Composite indicators using DEA

Murias *et al.* (2008) apply the DEA approach on data from 43 Spanish public universities in order

[29] For a more detailed discussion of weight restrictions and their effects on efficiency, see Allen *et al.* (1997).

[30] See Box A5 in Appendix A for a more technical exposition of composite indicators with DEA.

[31] For a more exhaustive list of applications, see Murias *et al.* (2008).

to build a composite indicator for quality assessment. The indicator comprises partial indicators – such as teaching staff to student ratio, number of exchange students, library seats per student, etc. – that aim to represent the most important facets and functions of university performance. A series of restrictions were introduced to the weights assigned to the various partial indicators. The quality index highlights the important differences in effectiveness among Spanish public universities.

2.4.1.3 Bootstrapping

The approaches we have covered so far do not include any estimation of the sampling distributions of the obtained efficiency scores. This implies that no information can be extracted on the confidence intervals for each efficiency estimator (i.e. a statistical estimate of how confident we can be in the estimate of efficiency that the model has generated). A significant and important contribution from Simar and Wilson (1998; 2000) shows how to use bootstrapping – a re-sampling technique – as a means of approximating the properties of the sampling distribution of the estimated efficiency scores and, hence, allowing one to conduct hypothesis tests and construct confidence intervals.

In its most simple form, the bootstrap involves randomly selecting thousands of 'pseudo samples' (using simple random sampling with replacement) from the available set of data. One then obtains 'pseudo estimates' from each of these samples. These thousands of pseudo estimates form an empirical distribution for the estimator of interest. This distribution is used as an approximation of the true underlying sampling distribution of the estimator.

Bootstrap methods became more widely used after sufficient computing power became available, but they are still a fairly involved exercise for the average applied researcher. In the case of DEA, bootstrap methods are further complicated by the one-sided nature of the inefficiency distribution, which generates bias and inconsistency problems in certain naive implementations of bootstrap methods. For an extensive discussion of these problems and some potential solutions, see Simar and Wilson (2000).

Box 13. Application: DEA with bootstrapping

Halkos and Tzeremes (2010) apply several DEA models on financial data of 23 Greek sectors relating to manufacturing, combining multiple financial measures in a single performance measure with the use of bootstrap techniques.

For applications in the context of education, see Box 14.

2.4.1.4 Malmquist productivity index

The Malmquist index is used to evaluate the productivity change of a DMU between two time periods. The concept was first introduced by Malmquist (1953) and further developed in the non-parametric framework by several authors[32]. This approach enables researchers to see the changes in productivity/efficiency across a number of DMUs between two periods of time – as well as look at changes within the two time periods. For example, researchers can look at the changes in efficiency between 2000 and 2010 and also study the annual changes in productivity[33].

Box 14. Application: Malmquist productivity index

Johnes (2008) derives Malmquist productivity indexes for 112 English HEI across the time period 1996/97 to 2004/05. Their findings suggest that Malmquist productivity has increased by an annual average of 1%. By further decomposing this increase into technical efficiency and

[32] See for example Färe and Grosskopf (1992), Färe *et al.* (1998) and Thrall (2000).

[33] For a technical exposition, see Box A6 in Appendix A.

technology changes, they reveal that it was driven by technology change – technical efficiency change was found to be negative.

Parteka and Wolszczak-Derlacz (2013) use bootstrap estimation procedures to obtain confidence intervals for the components of Malmquist indices of the productivity of 266 public HEIs in seven European countries over the period 2001–2005. They find considerable national differences, with HEIs in Germany, Italy and Switzerland performing better in terms of productivity change than HEIs from the rest of the countries in the sample.

Edvardsen *et al.* (2017) calculate a bootstrapped Malmquist productivity change index for HEIs in Norway, over the 10-year period 2004–2013. The main result of their analysis is the majority of the Norwegian institutions have had a positive productivity growth.

2.4.1.5 Two-stage Data Envelopment Analysis

All of the DEA models we have examined so far treat the production process as a 'black box' – that is, they consider the production process as a conversion of inputs into outputs, but without modelling how this conversion occurs. When DMUs are known to comprise several inter-related components (see Box 15 below), this knowledge is typically not accounted for in the modelling process. Either each component is evaluated in isolation from the others, or the analysis is conducted on the aggregate, without any consideration whatsoever of the way in which the component parts interact. In such a scenario, not all the available information is taken into consideration.

An alternative approach, originally proposed by Färe and Grosskopf (1992) and refined by Tone and Tsutsui (2009), is to construct a network of sub stages or nodes within the unit. Each node has inputs and outputs, with some outputs of some nodes acting as inputs into other nodes. These particular outputs are referred to as intermediate outputs or goods. The overall efficiency of the unit as a whole depends on the efficiency with which each node performs its role. In the case when the production process is modelled to consist of two nodes, a two-stage DEA analysis is applied[34].

Box 15. Application: Two-stage DEA

A notable application of two-stage DEA analysis in the context of HE in England is found in Johnes (2013). The author considers the case of one HEI within which there are two nodes. The first node 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 the second node, from which the ultimate outputs of the system are employability and student satisfaction. The data used refer to the academic year 2011/12 and the results indicate that, while institutions are typically highly efficient in converting inputs into outputs such as 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.

2.4.1.6 Free disposal hull

The free disposal hull (FDH) approach, proposed by Deprins *et al.* (1984), is a more general version of the DEA, as it relaxes a key assumption needed for basic DEA models (see technical extension Box 17 below). Free disposability means that inputs and outputs can freely be disposed of i.e., we can always produce fewer outputs with more inputs. Kerstens *et al.* (1994) pointed out that the minimal technical and behavioural assumptions underlying the FDH approach make it a particularly useful tool for analysing public sector efficiency questions[35].

[34] See Halkos *et al.* (2014) for an extensive discussion on two-stage DEA models.

[35] For a more technical exposition, see Box A7 in Appendix A.

Box 16. Application: Free disposal hull

Haelermans and De Witte (2012) use this approach to measure the efficiency in 119 Dutch secondary schools without imposing any *a priori* specification on the functional form of the educational production technology. Also, their model follows the order-m technique of Cazals *et al.* (2002), which mitigates the influence of outlying observations. The results suggest that educational innovation is positively related to efficiency. Profiling, pedagogic process and education chain innovations are all found to be significantly related to school efficiency.

2.4.2 Parametric frontier methodologies

In comparison to non-parametric frontier approaches, parametric frontier methods are more demanding with respect to assumptions, as they require the analyst to guess the shape of the frontier in advance by specifying a particular function relating the level of inputs to outputs. For example, whether the relationship is a straight line or a particular curve.

Box 17. Technical extension: Parametric frontier methods

In parametric approaches, we assume *a priori* that the production (or cost) function has a specific functional form:

$$f(x) = f(x_i; \beta)$$

The details of this function as defined by the parameters β are unknown. As with non-parametric approaches, we use actual data (on inputs x and outputs y) from various units to estimate the production function (that is, the β). We denote the estimated values by $\hat{\beta}$. Then, the estimated function is used to assess the performance of each unit.

A major difference between the two approaches is the estimation principle. Non-parametric methodologies rely on the idea of minimal extrapolation, while the parametric approaches use classical statistical principles, most notably the maximum likelihood principle. That is, we choose the value of $\hat{\beta}$ that makes the actual observations as likely as possible.

2.4.2.1 Stochastic frontier analysis

SFA is an alternative methodology for estimating the efficiency of DMUs. A researcher makes an assumption about how inputs are converted to outputs (e.g. by estimating the shape of the production function). The model is then estimated using econometric/statistical techniques which show what the optimal efficiency goal is and how far a DMU is from that goal (i.e. how inefficient the organisation is). SFA is more computationally demanding than DEA.

The literature on SFA, notably Debreu (1951), Farrell (1957) and Aigner *et al*. (1977), states that output of a DMU is a function of three components: i) the output made from a perfectly efficient use of inputs (i.e. how inputs would be used if a DMU was 100% efficient), ii) a random element often called ‘noise’ which describes unexpected occurrences or individual specific factors, and iii) the inefficiency of that DMU referred to as ‘technical inefficiency’.

Meeusen and van den Broeck (1977) extended the frontier approach introduced by Aigner and Chu (1968) in order to account not only for technical inefficiency, but also for measurement errors or statistical noise. Thus, they developed a statistical and theoretical method for measuring efficiency

that allows random events to contribute to variations in output[36].

Box 18. Application: SFA

Pereira and Moreira (2007) worked with data from 502 Portuguese secondary schools in order to examine their technical efficiency as well as the factors that determine the educational output. Interestingly, one of their results suggests that 'quality' of teachers has a greater impact on output than 'quantity'.

2.4.2.2 Two-stage SFA

A two-stage SFA approach involves adding another stage to the SFA estimation procedure. The second stage looks at the factors that could influence the efficiency or inefficiency and uses analysis to estimate the relative importance of these factors in determining how efficient or inefficient a DMU is[37].

Box 19. Application: Two-stage SFA

Scippacercola and D'Ambra (2014) use data on 35 secondary schools in the Campania region of Italy for the school year 2012/13 and the Tobit model to investigate the possible causes of technical inefficiency for various environmental variables, such as the cultural level of the students, the number of dropouts, parental involvement and school size. They show that only the cultural level of the students had a significant – and negative – effect on inefficiency scores.

2.4.2.3 SFA with panel data

Panel data record information about the same units over several time periods. For instance, a panel of stock market data could include the number of stocks and their prices over several days or weeks or months. In the HE context, a panel data set could include admissions data, funding availability, student teacher ratios, number of publications etc. over a number of academic years.

The repeated observations of the same unit over time allow researchers to capture individual specific effects that are not observable in the data. For example, stock market prices may depend on the companies' culture, which is not directly measurable. The appropriate type of panel SFA model to use depends on whether the inefficiency of organisations is believed to vary over time.

The concept of what the model is estimating in terms of the components described earlier is the same, the difference is that the models also deal specifically with changes over time.

If we believe that the inefficiency of a DMU is fixed over time, a researcher could use fixed or random effects SFA. If researchers believe that inefficiency of a DMU changes over time, they could use true fixed or random effects models or random parameters SFA.

[36] See Box A8 in Appendix A for a technical exposition of SFA and related issues.

[37] For a technical note on two-stage SFA, see Box A9 in Appendix A.

Box 20. Technical extension: Panel data SFA with constant inefficiency

Fixed effects SFA

A fixed effects SFA model allows researchers to account for unobserved differences among units or groups. The technical inefficiency estimates are allowed to be correlated with the independent variables of the model. Furthermore, as this model involves the estimation of the intercept for each unit to obtain the estimated values of efficiency, it is implicitly assumed that the most efficient unit is on the production frontier and the least efficient unit is normalised to zero. In order to yield a measure of technical inefficiency, the minimum estimated intercept is subtracted from the estimated intercept for each of the units for unobserved time invariant heterogeneity between units or groups. This suggests that the technical inefficiency estimates may be sensitive to including or excluding outliers with very low or very high values of the intercept. As with the standard fixed effects model, a fixed effect SFA model cannot accommodate time-invariant independent variables. Finally, many authors have shown that this model tends to overestimate efficiency (e.g. Greene, 2005).

Random effects SFA

Similar to a fixed effects model, a random-effects model also takes into account unobserved differences across units but assumes these are uncorrelated to the model's independent variables. Contrary to the fixed effects model, time-invariant independent variables can be included in a random effects SFA model. Such a time-invariant independent variable might for example capture whether an HEI is considered a traditional/old institution or if it receives public funding. Also, as shown by Greene (2005) and Farsi *et al.* (2005), this model tends to underestimate inefficiency.

Box 21. Application: Panel data SFA

Johnes and Johnes (2009) examine the cost efficiency of 121 English HEIs, over the period 2000-2003. In order to account for the fact that the institutional types in the sample are diverse, they use two models that allow cost function coefficients to vary across observations: a random effects model and a random parameters model.

Their results suggest that science undergraduates are costlier to produce in comparison to non-science undergraduates, and that postgraduate education is more costly than undergraduate education.

Box 22. Technical extension: Panel data SFA with inefficiency changing over time

True fixed effects SFA

Originally developed by Greene (2005), the true fixed effects SFA model allows for estimates of inefficiency to vary across time. The model includes a time invariant error term – that captures the unobserved heterogeneity among units – a random error term and a unit specific inefficiency term. This feature of the true fixed effects model enables the researchers to separate the effects into unobserved time-invariant and time-varying efficiency components.

*True random effects SFA*Also developed by Greene (2005), the true random effects SFA model differs from the original random effects SFA model by allowing the inefficiency term to vary across time. The main difference from the true fixed effects SFA model is found in the fact that the time invariant error term is uncorrelated to the model's other terms.

Random parameters SFA

Originally proposed by Hildreth and Houck (1968) and further developed by Tsionas (2002) and Greene (2005), the random parameters SFA model can be considered as a generalisation of Greene's true random effects SFA model. In addition to the intercept, other parameters in the model are also allowed to vary, and this reflects heterogeneity in the unit of analysis. As pointed out by Greene (2005), the parameters of this model cannot be estimated by traditional maximum likelihood techniques but require the use of the technique of maximum simulated likelihood[38]. However, this makes the model more computationally intensive than the approaches presented above.

Box 23. Application: Random parameters SFA

In the context of HE, random parameters SFA has been used by Johnes and Schwarzenberger (2011) to study the cost efficiency of HEIs in Germany. The authors find notable differences in the efficiency across HEIs, with small and specialised ones tending to be less efficient than others.

Agasisti and Johnes (2015) examine 954 public and private degree-granting HEIs in the U.S. over the period 2003-2006 and find that those with a high profile as research institutions tend to operate at high levels of efficiency.

For another application, see also Johnes and Johnes (2009) (Box 21).

2.4.2.4 Latent class SFA

A latent class is an unidentified subgroup, with similar unobservable characteristics, within a sample or a population. As a means of relaxing the assumption that the production technology is the same across units, Greene (2002) and Orea and Kumbhakar (2004) proposed an approach that divides the sample observations into such subgroups. While it may be desirable to allow DMUs to possess different production technology, this approach requires the researcher to choose in advance how many latent classes to include in the model (i.e. how many groups of DMUs share the same production technology), which adds another assumption to the modelling process. So far, no consensus has been established on what criteria this choice should be based on and some researchers start with the maximum number of latent classes that the statistical software can handle (see for instance Besstremyannaya, 2011)[39].

Box 24. Application: Latent class SFA models

Agasisti and Johnes (2015) adopt a latent class model – that incorporates a structural difference between two subsectors of the HE market – and a random parameters model. They note that:

'The distribution of efficiency scores is particularly wide in the case of the traditional frontier model, suggesting that when considering a unique frontier many colleges' results are deemed to be highly inefficient… Moreover, the number of highly efficient institutions is very limited. The picture dramatically changes when we turn to the latent class specification. Allowing each college to compare its efficiency with the frontier for its own group markedly improves the overall measured performance of institutions.'

Their findings further suggest that the differences between the two latent classes could be due to different scales of operation (number of bachelor and postgraduate students).

[38] See Greene (2001).

[39] For a technical note on latent class models, see Box A10 in Appendix A.

2.4.2.5 Bayesian SFA

All the SFA methods discussed previously are in the class of estimation approaches known as 'frequentist'. The alternative to a frequentist approach is a Bayesian approach[40] and this can also be applied to SFA, as for example in Mutz *et al.* (2017)[41].

In practical terms, while there are fundamental philosophical differences between the two approaches as well as different technical challenges in their implementation, they can both yield very similar results, and it is often straightforward to express a frequentist paradigm as a Bayesian one, and vice versa. That said, two practical observations are worth noting:

- Frequentist approaches are well understood and applied by a much larger body of researchers compared to Bayesian ones, and there is a much larger literature covering frequentist approaches
- In the context of big data that is frequently updated, Bayesian methods are a more 'natural' way to take into account newly available information – though again this process can be approximated by frequentist methods as well. Bayesian methods have been gaining in popularity due to their attractive properties in big data applications.

Box 25. Application: Bayesian SFA

Mutz *et al.* (2017) examine the efficiency of more than a thousand research projects funded by the Austrian Science Fund using Bayesian SFA. The study finds projects run by younger principal researchers and projects with longer durations have higher technical efficiency in comparison to projects run by older principal researchers and projects with shorter durations.

2.4.2.6 Semi-parametric DEA

While DEA is a non-parametric technique, it can be combined with regression analysis to create a semi-parametric DEA approach. The overall approach is similar to two-stage SFA. Typically, it involves a two-stage procedure: in the first stage, DEA efficiencies are evaluated and in the second stage the factors that explain the efficiencies are estimated using regression analysis. Usually, the factors in the second stage of the approach are different from the ones used in the first stage and are factors outside the control of the DMUs.

Semi-parametric DEA is for those who want to be able to determine which factors may influence the efficiency of DMUs, as well as identifying how efficient they are.

Box 26. Application: Semi-parametric DEA

Wolszczak-Derlacz (2017) examines the technical efficiency of public European and American HEIs across the time period 2000–2010. The efficiency scores are determined using DEA with different input-output sets and considering different frontiers: global frontier (all HEIs pooled together), regional frontier (Europe and the U.S. having their own frontiers) and country-specific ones. The results show that on average, traditional European HEIs are more efficient than their non-traditional counterparts, but this is not confirmed for American HEIs. Moreover, government funding seems to have a negative effect on the efficiency of universities in Europe, which again is not found in U.S. HEIs.

[40] A full discussion about the distinctions between Bayesian and frequentist approaches is beyond the scope of the report. Informal discussions can be found http://jakevdp.github.io/blog/2014/03/11/frequentism-and-bayesianism-a-practical-intro/ and http://andrewgelman.com/2018/03/24/economist-wrote-asking-make-sense-fit-bayesian-hierarchical-models-instead-frequentist-random-effects/

[41] The Bayesian approach was originally introduced to SFA by van den Broeck *et al.* (1994).

The semi-parametric DEA approach has also been used on secondary education and further education data. See for example Ray (1991), Lovell *et al*. (1994), Duncombe *et al*. (1997), Kirjavainen and Loikkanen (1998), Bradley *et al*. (2001; 2010) and Grosskopf and Moutray (2001).

Other examples of this approach can also be found in the literature. Oliveira and Santos (2005) work with data on Portuguese secondary schools and obtain FDH efficiency scores and slacks. In the second stage of their analysis, they apply the bootstrapping algorithm suggested by Simar and Wilson (2007), and construct confidence intervals for the parameters of various environmental variables. Their results suggest that unemployment rate, access to health care services, adult education and living infrastructures are in fact determinants of school efficiency.

3. Assessment of methods

3.1 Assessment strategy

Our plan for the assessment is to consider how each benchmarking methodology performs against a set of criteria shown in Table 2. The criteria have been selected to guide our discussion on factors to consider in selecting an appropriate methodology for benchmarking.

We have made the decision not to use numerical scoring but instead present a discussion about the advantages and disadvantages of each method for every criterion. We will not attempt to assign weights to any particular criterion as any weights will be both dependent on the precise situations in which benchmarks will be used and also down to the judgement of the analysts selecting the methodologies.

Consequently, the assessment does *not* result in a ranking of methodologies. We feel that without looking at the data for each specific context in which benchmarks are used, it would be too much of a stretch to produce a definitive ranking. In practice, all methodologies considered have their purpose and there will be circumstances in which every method is useful. Instead of numerical scores and rankings, the discussion will allow readers to understand the relative strengths and weaknesses of each benchmarking method. From this, analysts can use the information to make the judgements required about the appropriateness of each approach in various specific HE contexts.

Our assessment also includes a discussion of applications in a *general* HE context (e.g. addressing issues relating to small sample sizes and extreme values). Our review does not consider specific data issues (e.g. which benchmarking factors ought to be included). The parallel project run by Professor Draper is explicitly examining how the existing methodology can be adapted to deal with various specific data issues.

Table 2. Assessment criteria

Criteria		Description
Resources	Expertise	This criterion relates to the complexity and the specialist knowledge required to use the methodology.
	Level of effort	The appraisal will include an assessment of the time required to collect the necessary data and calculate the benchmark(s).
Communication	Interpretability	Interpretability refers to the extent a method allows us to tell a clear and insightful 'story'. Users of HE benchmarks include a wide range of stakeholders therefore being able to explain the methodology and results in plain English is an important consideration. This criterion relates to how intuitive and easily presentable the approach is, how simple it is to explain the results as well as whether it is possible to identify driving factors.
	Transparency	Transparency is about understanding how the benchmarks are calculated. When an approach builds on clear and transparent assumptions and the required calculations can be easily comprehended and/or replicated, it reassures stakeholders that the generated benchmark is credible.
Application	Scope	This criterion refers to the 'question' the benchmark attempts to answer, and how it is intended to be used. It may for example consist of a target that the HE providers should aim for or a threshold that their performances are compared to (e.g. where performance above the threshold is defined as excellent). The benchmark could also be a metric used for comparisons across all HE providers or for comparison of each HE provider with its peers.
	Technical issues	Given the characteristics of HE data, there are a number of technical issues of particular interest to the OfS, including whether the methodology can accommodate small samples, self-benchmarking, outliers in the data and whether it is vulnerable to excessive inclusion of benchmarking factors.

Box 27. Note on effect of sample size, outliers, self-benchmarking, and excessive inclusion of benchmarking factors

In developing any benchmarking application, similarly to most social science work, analysts need to operate with data generated and collected outside a pure experimental setting. The implication is that data often have 'problematic' features (missing values, measurement error, sampling variation) and – perhaps even more importantly – the explicit or implicit theory/model we seek to inform with the data is already known to at best be a rough approximation of the 'true' process under examination. In the words of George Box (1979), *'all models are wrong but some are useful'*.

The important takeaway for the purposes of our discussion is that none of the issues examined here (i.e. sample size, outliers, self-benchmarking, excessive inclusion factors) constitutes a fundamental 'problem' as far as statistical theory – which applies to idealised settings – is concerned. The uncertainty associated with small sample sizes can be fully quantified and outliers represent valuable information rather than a potentially problematic feature to be corrected. The right benchmarking factors to be included are stipulated by theory which can be tested and refined as needed.

The reason 'corrective measures' may be needed in practical applications is that real world data stray from the statistical ideal – and while in many cases the analyst will have a good idea of the possible issues, it is rare that these can be accurately quantified and formally modelled. Hence, and regardless of the method employed, analysts need to make practical decisions in the course of developing a benchmarking application to ensure their models (whether implicit or explicit) and

outputs are 'well behaved' – in the sense that they provide insights that can be confidently put to practical use.

The point to highlight is that the form these corrective measures should take, while grounded in and informed by statistical theory, can rarely be fully based on a thorough, formal analysis. Given that any particular choice cannot be formally justified, well-intentioned disagreements between experienced analysts are common, at least as far as the details of a particular approach are concerned, if not their basic motivation.

A comprehensive, formal treatment of issues relating to sample size, outliers, self-benchmarking and excessive inclusion of benchmarking factors would span several papers, and is thus beyond the scope of this text. Nevertheless, and without providing comprehensive justification or a formal treatment, it may be useful to highlight some important points and general guidelines.

A critical point to keep in mind is that if the theory/model (whether implicit or explicit) being applied is appropriate (i.e. certain underlying assumptions are met), then even with a small sample size we would expect our estimates to be centred around their 'true value' (in statistical terminology terms, our estimators are unbiased). In this context, sample size only affects 'sampling variation': any deviation of model estimates from the 'true values' will only be due to 'luck of the draw' rather than a deficiency of the model itself.

The flip side of this is that even a very large sample will not lead to good estimates if key model assumptions are not met. To give a simple example, if we are attempting to measure average student attainment and we have access to data on 90% of students, our estimates will still deviate from the 'truth' if the students with the lowest attainment are disproportionately likely to be amongst those not included in the data.

In summary, assuming the model being applied is appropriate, sample size considerations are only relevant with respect to sampling variation – the 'luck of the draw' effect. For example, if we compare two classes with similar students and find that class A recorded better results than class B, we can say confidently say that class A is *likely* better than class B – but if the samples being compared are small we will not be very confident this is not merely a result of 'luck of the draw' and hence would not be replicated if larger samples were compared.

Hence, in cases where the sample is not 'large enough', it may be desirable to develop a simpler model (at increased risk of excluding relevant factors) or to choose to suppress analytical results given a high probability they may be far from 'true' values. The best course of action is determined by practical considerations relating to the specific application, and there is no universally 'correct' answer based on statistical theory.

The discussion above refers to a sample being 'large' or 'large enough' loosely. Determining what is meant by a 'large' sample is not straightforward, especially *ex ante*: it depends on several factors including the variance of the process being considered, the desired accuracy of results, etc., and hence there is no simple answer to the question of 'minimum' sample size required in any given setting. That said, what follows is a discussion of some keys points to keep in mind.

There are diminishing returns in terms of model accuracy as the sample grows larger – there are large gains going from a sample of 10 to a sample of 100, while the gains in going from a sample of 100 to 200 are much smaller.

The more variability there is in the process, and the finer the effects/differences being modelled, the larger the sample size required.

A common misunderstanding is that sample size issues only arise due to the fact an analyst may only have access to a sample of the 'population' under examination (e.g. because data is only collected from a subset of all actual students or HEIs). When we attempt to go beyond simple descriptive statistics, however, in order to draw more fundamental conclusions ('causal inference'), the relevant 'population' is not the population of *actual* students or HEIs, but rather *all possible realisations* of those groups (i.e. the 'population' is infinite). If we have data on the entire population of students we can confidently say that HEI A performed better than HEI B *in the specific instance*, but this does not necessarily translate to HEI A being a *better performer* than HEI B. The particular outcomes observed could well have been down to 'luck of the draw' rather

than fundamental differences between the two HEIs.

Sample size is intimately related to the other technical issues discussed here. With a large enough sample, there is less chance that an outlier merely arose due to 'luck of the draw', would unduly affect results and should thus be excluded from estimations. An outlier in the context of a large sample should generally be treated as yet another observation, naturally arising in the context of the process being considered and carrying valuable information. In the context of a 'small' sample, the presence of an outlier increases fears that the particular sample being considered may be unrepresentative – so as a practical matter an analyst may choose to 'weigh it down' or even fully exclude it from estimations.

Similarly, with a 'large' sample, self-benchmarking will only have a minimal impact on results, and any irrelevant benchmarking factors would either be assigned a weight of zero (in stochastic methods) or be easily identified as irrelevant without affecting the results when deterministic approaches are used. To give an example, if a variable for brown-eyed students was included, in a small sample there is the real possibility that merely by 'luck of the draw' – in terms of a particular cohort of students – eye colour may be found to affect attainment. In a large enough sample, a stochastic method would end up applying a zero weight to 'brown eyes', while a deterministic method would yield identical results for brown-eyed and other students, and hence have no impact on headline results.

Hence, with small sample sizes, there is a more pressing need to exclude irrelevant factors based on theory/prior knowledge, as there is a greater probability they will lead to 'overfitting' the model – estimating a relationship that is a mere artefact of the specific sample and would not be replicated if a different sample was obtained. Again, no simple answers are at hand. The analyst needs to utilise her experience and prior knowledge to assess the likely benefits and costs of including a factor that may or may not be relevant.

Given that in most cases it is not possible to formally determine the best course of action with respect to outliers, sample size cut-off points, etc., sensitivity analysis and experimentation can play a key role. If key results do not change much regardless of formulation (i.e. results are robust to different modelling choices), we can be more confident they provide constructive insights. If fundamental differences emerge, a more in-depth investigation is called for.

Box 28. Benchmark gaming

Even though it is common for practitioners to refer to 'benchmark gaming' (Hazelkorn, 2011; Department for Education, 2017), we found no agreed formal definition in the literature. However, it is widely accepted that gaming has taken place when a provider takes conscious actions that result in improving the figures capturing performance, without any genuine underlying improvements occurring. Such activities might for example include reconfiguring the underlying factors (e.g. student recruitment), making courses easier and manipulating or even misreporting data[42].

The reasons behind benchmark gaming are quite straightforward. By gaming the benchmark, a provider appears to be performing better in comparison to its peers. For example, in the case of university rankings, gaming can lead to rapid improvement in the provider's relative position. The provider is then more likely to gain visibility and improve its academic reputation, attract more talented staff, more and academically stronger students and possibly charge higher tuition fees[43].

Intuitively, 'relatively' simple methodologies are considered easier to game than more complex ones. Yet, as such approaches tend to be more transparent and easier to interpret, determining whether gaming has taken place could also prove to be more straightforward. On the other hand, the use of a technically complex benchmarking approach would make it harder for the institutions to game the benchmark, but it could also be more difficult to identify and control for gaming.

[42] See for instance Bhattacharjee (2011) and Pérez-Peña and Slotnik (2012).

[43] See Monks and Ehrenberg (1999) and Meredith (2004) for a more extensive discussion.

3.2 Assessment

3.2.1 Deterministic methodologies

Resources: Expertise and level of effort

Class 1 and Class 2 include the least technically complex benchmarking methods in use. They both involve simple calculations that do not necessarily require advanced mathematical or statistical skills, and the benchmarks can be generated and recalculated with minimal effort. On the other hand, Class 3 methods involve more demanding calculations, especially when we consider that indirect or direct adjustments are usually made (see Section 2.1.2 for more details), and generally require the contribution of experienced analysts. Furthermore, as a single benchmark has to be estimated for every unit under examination, the calculations have to be executed multiple times. As a consequence, generating and recalculating the benchmarks is more time consuming.

Communication: Interpretability

As Class 1 and Class 2 methods are among the simplest ones available, it is often easy for users to develop an intuitive understanding of how everything has been put together and how different elements affect headline findings. This does not suggest that no complexities are present - for example, weights could be determined in a complicated manner, or the indices used may not be standard natural units. However, the general frameworks utilised tend to be amongst the easiest to understand and interpret.

In contrast, Class 3 methods can be much more difficult to explain and understand, and specialist knowledge and significant effort is often required to grasp how the general framework works and what drives the results.

Communication: Transparency

Class 1 and Class 2 methodologies are generally straightforward to fully document, and even relatively inexperienced analysts will usually be able to understand the various steps involved and replicate them. Class 3 methodologies can also be fully documented, but the higher level of complexity can complicate matters significantly. Replicating the results should generally be feasible in principle, but the level of effort and expertise required to do so often renders this difficult in practice.

Often Class 1 and Class 2 methodologies are more frequently cited. Some of the more advanced methods, including Class 3 approaches, require specialist knowledge to allow users to develop an intuitive understanding of how they work, and it can be difficult to explain results and their limitations to non-technical audiences.

Box 29. A note of caution on transparency and interpretability

The fact that Class 1 and Class 2 approaches are in principle transparent and easy to interpret does not necessarily imply that users invariably develop a correct understanding of either how they work or their limitations. In fact, users without a solid statistical background may often *feel* they have an intuitive understanding, when in fact they misunderstand key aspects of the method and results.

This is a critical point to keep in mind, especially given these methods are widely cited.

Box 30. University rankings: a discussion

This box revisits some common criticisms of Class 1 methods, often expressed in the context of university rankings – a sub-category of applications that fall into this class. Turner (2005) states that: *'(the process) involves adding indicators which have completely different scales, and the variations of which are not comparable.'*

While this criticism is often waged at university rankings, it is not a fundamental feature of Class 1 methods – in fact, it is possible to normalise indicators' scales, look at variations which are comparable or even adapt the data to meet any properties the researcher believes they should meet. Turner (2005) also states that:

'Diversity of mission cannot be recognised in a single league table of this kind. Institutions that specialise in research or in teaching will necessarily fall down the table because they have low scores in one of the important measures.'

Even though this is a valid point for various well-known university rankings, it is more related to their specific aim and motivation rather than a fundamental feature of Class 1 approaches. As far as the general method is concerned, there is nothing stopping the researcher restricting comparisons amongst specific subgroups of HEIs that have a common mission, or for that matter to apply weights in a way that does not penalise HEIs which focus on either research or teaching vis-à-vis universities that focus on both.

Finally, it is worth noting that there are instances in the literature where university rankings are found to rank institutions in a way that is very similar to rankings produced by more complex approaches. For example, Sarrico *et al.* (1997) use DEA to assess the desirability of English HEIs to various types of prospective students. Their results suggest that for the academically strongest students (i.e. students who are not restricted in their choice of university by entry requirements), DEA generated rankings which were almost identical to the rankings reported in the Times Good University Guide.

All the above indicate that well-crafted rankings may indeed carry useful information which they communicate efficiently to their users, without exposing them to details of little relevance to them (e.g. whether an HEI is over or under performing given its level of funding or other characteristics).

Application: Scope

Class 1 benchmarks are widely used by prospective students and employers. Perhaps largely due to the wide publicity they receive, they are also used by HEIs themselves for marketing and other purposes[44]. Our impression is that users do not actually see them as tools that capture the potential for improvement or as tools to uncover areas of inefficiency – or at least realise that they are not generally built for this purpose. It is worth noting that the method may be adapted to cover these elements (e.g. by taking into account spending per student) but such analysis is not done in a formal input-output framework – as tends to be the case with more complex methods. Another feature of Class 1 approaches is that they generally tend to allow for global as well as local comparisons.

Compared to Class 1 approaches, Class 2 methodologies exhibit enhanced flexibility with regards to the questions they can provide answers to. For example, calculating an external benchmark often enables the researcher to consider specific conditions. It is possible to estimate a benchmark based on information on all HEIs in the sector – appropriate for global comparisons – or based on selected peers – appropriate for more local comparisons. A researcher may also apply explicit controls in

[44] A non-exhaustive list of such HEIs includes: Cardiff University (http://www.cardiff.ac.uk/news/view/937922-welsh-university-of-the-year-2018), Newcastle University (https://www.ncl.ac.uk/about/quality/league-tables/#d.en.251551), University of Bristol (http://www.bris.ac.uk/chemistry/courses/undergraduate/league-tables.html), University of Exeter (https://www.exeter.ac.uk/about/facts/success/), University of St. Andrews (https://news.st-andrews.ac.uk/archive/st-andrews-top-in-the-uk-for-student-experience/), University of Sheffield (https://www.sheffield.ac.uk/about/rankings#_ga=2.62313772.1856828852.1530599713-125457288.1530599713), University of Surrey (https://www.surrey.ac.uk/about/facts/rankings-league-tables), and University of Warwick (https://warwick.ac.uk/#Rankings/).

terms of measuring against what a particular HEI is expected to achieve given its underlying characteristics (e.g. prior attainment of its students, its budget etc.).

Class 3 benchmarks tend to be used by government officials, HEIs themselves, and to a lesser extent, by prospective students and employers. They succeed in capturing the fact that performance comparisons across all HE providers might not be meaningful, for example due to differences in their mission or the resources they have to work with. They are geared towards the identification of areas where efficiency improvements may be possible but tend to not allow direct comparison between relatively dissimilar institutions.

Application: Technical issues

In all classes of deterministic benchmarking methodologies, the appropriate treatment of small samples is a matter of design choice and transparency. Their inclusion should be examined and if the results are found to differ substantially under different design choices, further examination is advisable.

As far as self-benchmarking is concerned, since in Class 1 approaches the performance of each provider is measured using information only from that provider, no issues arise. In Class 2 approaches, including the reference HEI in the estimation is highly likely to affect the benchmark. If for example, this HEI performs relatively better than the rest of the institutions, its inclusion would raise the average and thus lead to an underestimate of the extent it outperforms the rest of the sector.

Class 3 benchmark estimates are commonly based on subclasses or subgroups, determined by qualitative judgement or cluster analysis[45]. Consequently, it is possible to end up generating benchmarks or indicators based on a small number of observations, resulting in misleading or difficult-to-interpret indicators. A common practice towards avoiding this is to require a minimum amount of observations (sample size cut-off point) when calculating benchmarks. For instance, the methodology currently applied by HESA excludes indicators in cases where an HE provider has fewer than 20 students with known data[46].

When outliers in the data are present, it is possible that Class 1 benchmarks may be affected, assuming the output is a composite indicator. That said, there are treatments that allow the researcher to correct for such irregularities, such as normalising to a particular scale (for example 100) or using an index. The benchmark obtained in Class 2 and Class 3 methods is also likely to be affected by outliers. In these approaches, however, it might be interesting to examine in what sense a provider is considered an outlier: on the input side (e.g. for having too many students with particular entry qualifications so there are no comparable peers) or output side (e.g. too many graduates with exceptional performance) etc. These insights may affect the choice of corrective measures to be applied.

3.2.2 Stochastic methodologies

Resources: Expertise and level of effort

Even though stochastic approaches usually involve highly technical calculations, the appropriate estimates can generally be obtained relatively quickly and easily with the use of modern statistical software. OLS is one of the simplest stochastic methods that can be used for benchmarking. It is relatively easy to implement and is familiar to most researchers as it is widely taught in quantitative research methods courses.

[45] See for example Erdoğmuş and Esen (2016) and Jatmiko *et al*. (2017).

[46] Class 1 and 2 approaches tend to be applied at aggregated data, so issues relating to small samples do not arise with the same frequency as is the case with Class 3 approaches.

Other stochastic approaches are generally more complicated and therefore require more specialist knowledge. Although they offer the option to accommodate various extensions (e.g. value-added HLM-based approaches allow for multilevel analysis) it may prove difficult and time consuming to identify where they should be used in practice and how to overcome any technical issues. For example, researchers working with panel data have to deal with issues such as unbalanced panels.

Communication: Interpretability and transparency

OLS-based approaches generate results that are straightforward to understand as they involve a simple calculation of the difference between a DMU's position and an estimated average efficiency level, conditional on the DMU's characteristics. The method itself is relatively easy to explain to non-experts and is intuitive to understand (as are COLS and MOLS). One minor difficulty is that it is less straightforward for an individual who does not have a technical background to understand the assumptions behind the OLS model and hence identify its limitations, let alone understand the limitations of more advanced methods. For example, OLS and HLM-based value-added models are built on the assumption that the entry test scores and current average test scores have a linear relationship. If this assumption is not met, the models generate biased estimates. Also, the reliability of the regression residuals is limited when two variables are highly (linearly) related.

A difficulty common to all stochastic approaches comes from the fact that the interpretation of regression residuals as a pure measure of efficiency can be easily challenged and thus may be confusing to some users. The main reason behind this, as explained previously, is that residuals typically capture not just inefficiency but also measurement errors, noise or even unobservable differences across DMUs.

Application: Scope

Due to their simplicity, OLS-based stochastic methods are widely used for benchmarking purposes. Yet, Barrow and Wagstaff (1989) point out that OLS regression residuals provide a measure of efficiency relative to 'average' performance, rather than the 'best practice' frontier. Hence, this measure sheds no light on how far each provider may be from the most efficient practices i.e. how efficient each provider is in comparison to the frontier (though the COLS extension does deal with this critique). Unless properly understood, the inexperienced user could misinterpret the results of the benchmarking analysis.

This issue – along with the questionable practice of interpreting the regression residuals as a pure measure of efficiency – led to the development of the frontier methodologies covered in Section 2.4 and their wide adoption in many benchmarking applications.

As far as flexibility is concerned, OLS-based approaches generate estimates appropriate for global comparisons, though it is also possible to conduct analysis on specific sub-groups in the sample. Some of the more complex approaches can generally account for more driving factors and are geared towards capturing issues of special interest.

For instance, panel data analysis is the tool of choice when there are important characteristics driving performance that do not change (much) over time and for which no useable data exists (e.g. an institution's reputation). However, since SFA panel data methods were designed specifically to measure efficiency they are more appropriate to use in a benchmarking context rather than straightforward panel data models (see Sections 2.4.2.3 and 2.4.2.4)[47].

Application: Technical issues

As is the case with deterministic approaches, design choice and transparency are important determinants of the appropriate treatment of small samples with stochastic methods as well. For

[47] Note however that SFA, either with cross-sectional or panel data, tends to be used with data at the organisation level, whereas other stochastic approaches tend to use individual level data.

example, considering a hierarchical structure while estimating value-added scores helps utilise all available information at the student level and can more accurately capture the relationship between current test scores and entry test scores. That said, when working with small samples fewer variables can be included in the analysis, and hence more complex stochastic approaches are more difficult to work with and likely to generate misleading results. This is also true when considering including benchmarking factors – a smaller sample would generally allow for fewer factors[48].

In other words, a larger sample size can help increase the reliability of residuals and thus lead to more trustworthy findings, and the more complex the method the more 'data-hungry' it tends to be. Hence, the use of a relatively simpler method might be preferable when the sample size is small.

As highlighted in Box 27, the presence of outliers in the data is not always a cause for concern. Yet, when extreme data points do not reflect genuine differences between DMUs – e.g. due to measurement error – they are likely to bias estimates away from their true values and corrective measures ought to be taken. OLS estimates, along with estimates from many other stochastic approaches, can be highly sensitive to such extreme values. If this is the case, robust regression methods could be used as an alternative to OLS. The appropriate treatment of outliers in the data is further considered in the following section, where frontier methodologies are assessed.

Box 31. Robust regression methods

The term 'robust estimation' does not refer to a single procedure. It is rather used to describe a whole class of techniques (e.g. least trimmed squares, Theil–Sen estimator, etc.) for assessing the validity of a set of empirical results and, ultimately, that of the underlying theoretical model. Essentially, these techniques attempt to dampen the influence of outlying cases in order to provide a better fit to the majority of the data. At the heart of all robust regression methods is an ability to apply corrective measures so that flaws in the data or the sampling process are to some extent accounted for.

3.2.3 Frontier methodologies

Resources: Expertise

Both parametric and non-parametric approaches involve a degree of mathematical complexity, with parametric methods generally being more technically demanding. The latter also require more assumptions to be made in order to produce estimates. For instance, they require the researcher to choose *a priori* the mathematical expression that captures the transformation of inputs into outputs. This is not true for non-parametric approaches and in fact FDH models tend to be even more flexible than the standard DEA models. On this subject, Pereira and Moreira (2007) point out that: *'the relationship between inputs and output in the educational process is rather complex and can only be summarised imperfectly in a production function.'*

Haelermans and De Witte (2012) note that because of this feature, non-parametric approaches are well-suited to applications in the context of education:

> *'This is convenient as information on the relationship between the resources and the produced outputs is often unavailable to researchers (e.g. Yatchew, 1998). As parametric models assume a priori a functional form on this relationship, they might be wrongly specified which leads to biased estimation results (Hjalmarsson et al., 1996).'*

It is noteworthy that if it is deemed important to model how the conversion of inputs into outputs occurs – and in particular if this process is conceived to comprise several interrelated components – the non-parametric framework can accommodate for such a design choice. Halkos *et al.* (2014)

[48] See also Box 27.

conclude that:

'...when there are more complex systems than a simple input–output procedure, [the standard single-stage DEA approach] fails to address the internal structures. A decision maker needs a tool which can incorporate these interrelations into the model and provide more accurate results in order to monitor the overall and individual procedures more effectively and make better decisions.'

If this is the case, a two-stage (or even multi-stage) DEA can be applied.

Furthermore, as parametric approaches accommodate both inefficiency and measurement errors, they require assumptions on their statistical properties as well. Even though some authors verify that efficiency estimates are relatively robust to the choice of distributional specifications of the inefficiency term (e.g. Greene, 1990; Ruggiero, 1999; Kumbhakar and Lovell, 2000; Scippacercola and D'Ambra, 2014), others find that different assumptions on the distribution of efficiency imply quite different estimates of efficiency (see for example Baccouche and Kouki, 2003). Considering the remark of Parmeter and Kumbhakar (2014):

'Most applied papers do not rigorously check differences in estimates and inference across different distributional assumptions…',

we may conclude that, even though the effect of distributional specifications of the error and inefficiency terms on efficiency estimates are not yet clear and further research remains to be done, it may prove wise to adopt relatively simple assumptions on their distribution (Ritter and Simar, 1997).

Resources: Level of effort

Since non-parametric approaches involve obtaining the level of efficiency for each unit separately, they require solving the same problem multiple times in order to obtain a benchmark for every unit under examination. Parametric models on the other hand give efficiency estimates for all units in a single step. However, nowadays both of these processes are automated, and the benchmarks are calculated with the use of specialised software. Depending on the complexity of the models adopted, the size of the dataset and the mathematical tools used to generate the estimates[49], the time required could vary substantially. For example, a sophisticated panel data SFA model is a computationally more intensive exercise than a relatively simpler DEA setting and is highly likely to take longer to produce efficiency estimates.

The two criteria discussed above clearly indicate that experienced and knowledgeable researchers are required to produce estimates with the use of frontier methodologies, but non-parametric approaches might prove easier to work with.

Communication: Interpretability and transparency

In essence, the level of efficiency in both parametric and non-parametric approaches is characterised by a single metric and thus comparisons between providers are usually straightforward. As far as the steps of the analyses are concerned, inexperienced users may struggle to follow the steps these methodologies involve – or the rationale for their inclusion – in particular when more advanced parametric approaches are considered (e.g. panel data models).

Presenting the main elements of each approach in simple terms, along with visual aids, can help audiences with more limited backgrounds in mathematics and statistics grasp how the efficiency estimates are obtained, along with their intended use – but even in this case, it is unlikely a reader with limited technical background will be able to fully grasp the rationale, limitations and appropriate use of the methods.

Also, since these approaches usually require advanced mathematical knowledge, access to a great

[49] Non-parametric approaches involve solving linear programmes, while parametric methods usually require more advanced mathematical techniques to produce estimates (e.g. maximum likelihood or maximum simulated likelihood).

amount of data and specialised software, it can prove quite challenging to replicate results.

Caution should be applied in clarifying the special features and limitations of each one of those methods. For instance, non-parametric models tend to ascribe all deviations from the frontier to inefficiency, because they neither account for measurement errors nor 'noise'. Parametric models on the other hand can accommodate both noise/measurement errors and inefficiency. Bradley *et al.* (2010) point out that: *'it is important for the efficiency measurement approach to be able to handle such errors'*, since some HEIs' outputs – student employability and the number of publications for example – could be greatly affected by random factors or measurement error.

As far as determining the factors that drive efficiency levels is concerned, either in parametric or non-parametric approaches, a strategy widely adopted in both settings is adding another stage to the analysis. In both cases, this step involves singling out the factors that could influence the efficiency estimates with the use of regression techniques (see Sections 2.4.2.2 and 2.4.2.6).

It should be noted however that even though this flavour of two-stage DEA is popular in the literature, McCarty and Yaisawarng (1993) observe that its use could be problematic when there is correlation between the inputs of the first stage and the independent variables of the second stage. In this case, the DEA scores computed in the first stage could be biased. Simar and Wilson (2011) *'do not recommend the use of second-stage regressions involving DEA efficiency scores'* because the method *'is consistent only under very peculiar and unusual assumptions […] that limit its applicability.'*

Finally, an important note before leaving this section is that the concept of efficiency is not tautological to improvement. As pointed out for example by Turner (2005):

'the score of 100 or an efficiency of 100% in DEA does not imply that an institution is incapable of improvement. A score of 100 simply means that an institution has achieved a place on the current data envelope.'

Box 32. Confidence intervals in non-parametric approaches

A notable drawback of non-parametric methodologies is that the associated statistics do not follow a known (or assumed) distribution, and hence it is not possible to directly assess the possible impact of sampling variation and estimate confidence intervals.

In order to provide a workaround to this problem, Simar and Wilson (1998) introduced bootstrap DEA, allowing for the sensitivity of efficiency scores which results from the distribution of efficiency in the sample to be assessed. The use of the bootstrapping technique is not exclusively limited to standard DEA but could also be applied to other deterministic approaches, including the deterministic methods covered above.

Application: Scope

The benchmark estimates that emerge from these methodologies aim to help answer the question of how efficiently each unit operates. Even though efficiency comparisons across providers are straightforward, they do not appear to be useful to non-technical audiences – possibly due to their specialised origin – but are commonly used among academics and experienced professionals with a solid mathematical or statistical background.

A notable exception could be the application of DEA to build single composite indicators, and thus produce a ranking of the units under examination. As noted by Murias *et al.* (2008):

'Single indicators are particularly useful because they facilitate the compilation of quantitative information regarding a given system. However, in order to summarize the characteristics of said system, there must be a common trend, which is not always easy to identify. One way of overcoming this problem is to use composite indicators, which constitute useful instruments for summarizing complex or multidimensional issues and, in aiding policy making and increasing public awareness.'

This approach has the merits of non-parametric frontier analysis while producing results in a way that can be easily understood even by people who have little understanding of statistical techniques.

Moreover, both approaches are flexible enough to accommodate for the application of techniques developed to enable the researcher to obtain improvement measures. The Malmquist productivity index in a non-parametric setting and panel data analysis in parametric methodologies are important examples. A notable limitation of the Malmquist productivity index is that it is built to capture improvement only between two time periods – but these time periods do not necessarily have to be consecutive. Johnes (2008) for instance estimates the improvement of English HEIs from 1996/97 to 2004/05. On the other hand, panel data SFA models can be used to estimate improvement across longer time periods, but they are more complex and require a great amount of data to generate accurate results.

Application: Technical issues

As with all other statistical exercises, outliers in the data should be investigated carefully. As noted by Coelli and Perelman (1999), Johnes and Johnes (2009), Halkos *et al.* (2014) and Scippacercola and D'Ambra (2014), outliers significantly affect the frontier generated by non-parametric approaches and thus efficiency rates. However, Pereira and Moreira (2007) claim that (in comparison to DEA and FDH) *'stochastic frontier analysis is […] less sensitive to the presence of outliers'.*

As discussed previously, deciding whether to include or exclude extreme observations from the analysis mainly depends on practical considerations relating to sample size, the aim of the exercise, as well as considerations of transparency. If the results obtained differ substantially based on how outliers are treated, further examination of the reasons why this occurs is required. In particular, if there is reason to suspect that such observations have arisen due to measurement error, they should be excluded from the dataset or have their effect on efficiency estimates mitigated.

With regards to sample size, the literature concludes that it has a significant influence on the performance of frontier approaches. To be more specific, Andor and Hesse (2011), motivated by Banker *et al.* (1993) and Ruggiero (1999), examine which factors affect the performance of DEA and SFA by applying Monte Carlo analysis[50]. Similar to Banker *et al.* (1993), the performance criterion used to make comparisons between the two approaches is the mean of the absolute deviation of true versus estimated efficiencies. Their findings are largely in accordance with the ones obtained by Banker *et al.* (1993) and suggest that for samples including 50 or more DMUs, SFA achieves better results than DEA, while DEA performs better for samples with fewer than 50 DMUs[51].

As far as excessive inclusion of benchmarking factors is concerned, Bradley *et al.* (2010) points out that: *'One well-known disadvantage of DEA is that the degree of discrimination between DMUs is lower the more variables are included, and so a parsimonious DEA model is to be preferred.'*

Moreover, as noted by Johnes and Johnes (2009), in non-parametric methodologies there is no statistical inference to help the analyst decide whether a variable should be included or not, as would be the case for stochastic techniques. Selection of variables in this case is done mainly on the grounds of theory, although insights may also be gleaned by applying sensitivity analysis.

[50] Andor and Hesse (2011) consider an output-orientated VRS DEA model, while for the SFA model they assume a Cobb-Douglas production function, a normal distribution for the noise term and a half-normal distribution for the inefficiency term.

[51] Contrary to Banker *et al.* (1993) however, Andor and Hesse (2011) show that SFA performs better than DEA for samples with fewer than 20 DMUs.

4. Conclusions

This report has presented an assessment of a broad range of benchmarking methodologies and described examples of their applications.

In the HE sector, benchmarking is used for a broad set of purposes. One application is around improving efficiency, however there are many others including providing information that prospective students can use to select their preferred HE provider. With such a range of benchmarking uses, providing 'tools not tables' may be advantageous. In other words, instead of publishing the results of a benchmarking exercise, it may be useful to develop platforms/approaches that allow users to generate their own benchmarks and analysis in a flexible way.

Benchmarking is not a purely academic exercise in the sense that there is no objectively optimal benchmark. Rather, the performance of a benchmark depends on its outputs and how its users understand and engage with the results. Assessing this performance will depend on the extent to which a benchmark has impact e.g. how much it has helped to drive up standards.

Consequently, the most appropriate benchmarking methodology may not be theoretically elegant or difficult. In the HE context with a wide set of users with varying degrees of statistical knowledge, a simple approach may be more useful and achieve a greater impact on improving the sector.

List of Abbreviations

CJG	Criminal Justice Group
COLS	Corrected ordinary least squares
CQC	Care Quality Commission
CRS	Constant returns to scale
DEA	Data envelopment analysis
DfE	Department for Education
DLHE	Destination of Leavers from Higher Education (survey)
DMU	Decision-making unit
EBacc	English Baccalaureate
ESMU	European Centre for Strategic Management of Universities
FDH	Free disposal hull
GCSE	General Certificate of Secondary Education
HE	Higher education
HEFCE	Higher Education Funding Council for England
HEI	Higher education institution
HESA	Higher Education Statistics Agency
HESPA	Higher Education Strategic Planners Association
HLM	Hierarchical linear model
HMPPS	Her Majesty's Prison and Probation Service
MOLS	Modified ordinary least squares
MQPL	Measuring the Quality of Prison Life
NACUBO	National Association of College and University Business Officers
NOMS	National Offender Management Service
NSS	National Student Survey
OfS	Office for Students
OLS	Ordinary least squares
PCF	Potential confounding factor

PRS	Prison Rating System
SFA	Stochastic frontier analysis
SHMI	Summary Hospital-level Mortality Indicator
TEF	Teaching Excellence and Student Outcomes Framework
VRS	Variable returns to scale

References

Abbott, M., Doucouliagos, C., 2003. The efficiency of Australian universities: a data envelopment analysis. Economics of Education Review 22, 89–97. https://doi.org/10.1016/S0272-7757(01)00068-1

Agasisti, T., Johnes, G., 2015. Efficiency, costs, rankings and heterogeneity: the case of US higher education. Studies in Higher Education 40, 60–82. https://doi.org/10.1080/03075079.2013.818644

Aigner, D., Chu, S.F., 1968. On estimating the industry production function. The American Economic Review 58, 826–839.

Aigner, D., Lovell, C.A.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. Journal of Econometrics 6, 21–37. https://doi.org/10.1016/0304-4076(77)90052-5

Aleskerov, F.T., Belousova, V.Y., Petrushchenko, V.V., 2017. Models of data envelopment analysis and stochastic frontier analysis in the efficiency assessment of universities. Autom Remote Control 78, 902–923. https://doi.org/10.1134/S0005117917050125

Allen, R., Athanassopoulos, A., Dyson, R.G., Thanassoulis, E., 1997. Weights restrictions and value judgements in data envelopment analysis: Evolution, development and future directions. Annals of Operations Research 73, 13–34. https://doi.org/10.1023/A:1018968909638

Andor, M., Hesse, F., 2011. A Monte Carlo simulation comparing DEA, SFA and two simple approaches to combine efficiency estimates. CAWM Discussion Paper No. 51, University of Munster.

Baccouche, R., Kouki, M., 2003. Stochastic production frontier and technical inefficiency: A sensitivity analysis. Econometric Reviews 22, 79–91. https://doi.org/10.1081/ETC-120017975

Baltagi, B., 2005. Econometric analysis of panel data, 3rd ed. John Wiley & Sons.

Banker, R.D., Charnes, A., Cooper, W.W., 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science 30, 1078–1092. https://doi.org/10.1287/mnsc.30.9.1078

Banker, R.D., Gadh, V.M., Gorr, W.L., 1993. A Monte Carlo comparison of two production frontier estimation methods: Corrected ordinary least squares and data envelopment analysis. European Journal of Operational Research 67, 332–343. https://doi.org/10.1016/0377-2217(93)90289-Y

Barrow, M., Wagstaff, A., 1989. Efficiency Measurement in the Public Sector: An Appraisal. Fiscal Studies 10, 72-97. https://doi.org/10.1111/j.1475-5890.1989.tb00339.x

BBC News (2011). “Gaming” the school league tables?, URL https://www.bbc.com/news/education-12914964

Besstremyannaya, G., 2011. Managerial performance and cost efficiency of Japanese local public hospitals: A latent class stochastic frontier model. Health Econ 20, 19–34. https://doi.org/10.1002/hec.1769

Bhattacharjee, Y., 2011. Saudi universities offer cash in exchange for academic prestige. Science 334, 1344–1345. https://doi.org/10.1126/science.334.6061.1344

Bradley, S., Johnes, G., Millington, J., 2001. The effect of competition on the efficiency of secondary schools in England. European Journal of Operational Research 135, 545–568. https://doi.org/10.1016/S0377-2217(00)00328-3

Bradley, S., Johnes, J., Little, A., 2010. Measurement and determinants of efficiency and productivity in the further education sector in England. Bulletin of Economic Research 62, 1–30. https://doi.org/10.1111/j.1467-8586.2009.00309.x

Box, G.E.P., 1979. Robustness in the strategy of scientific model building, in: Robustness in statistics. Academic Press, pp. 201–236. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Care Quality Commission, 2018a. How CQC monitors, inspects and regulates NHS trusts. March 2018.

Care Quality Commission, 2018b. Key lines of enquiry, prompts and ratings characteristics for healthcare services. March 2018.

Cazals, C., Florens, J.-P., Simar, L., 2002. Nonparametric frontier estimation: A robust approach. Journal of Econometrics 106, 1–25. https://doi.org/10.1016/S0304-4076(01)00080-X

Charnes, A., Cooper, W.W., 1962. Programming with linear fractional functionals. Naval Research Logistics Quarterly 9, 181–186. https://doi.org/10.1002/nav.3800090303

Charnes, A., Cooper, W.W., Golany, B., Seiford, L., Stutz, J., 1985. Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. Journal of Econometrics 30, 91–107. https://doi.org/10.1016/0304-4076(85)90133-2

Charnes, A., Cooper, W.W., Rhodes, E., 1979. Short communication: Measuring the efficiency of decision-making units. European Journal of Operational Research 3, 339–338.

Charnes, A., Cooper, W.W., Rhodes, E., 1978. Measuring the efficiency of decision making units. European Journal of Operational Research 2, 429–444. https://doi.org/10.1016/0377-2217(78)90 90138-8

Coelli, T., Rao, D.S.P., O'Donnell, C.J., Battese, G.E., 2005. An Introduction to efficiency and productivity analysis, 2nd ed. Springer US.

Coelli, T., Perelman, S., 1999. A comparison of parametric and non-parametric distance functions: With application to European railways. European Journal of Operational Research 117, 326–339. https://doi.org/10.1016/S0377-2217(98)00271-9

Colledge, L., 2017. Snowball metrics recipe book, 3rd ed. Elsevier.

Complete University Guide, 2018. University and Subject League Tables Methodology. URL https://www.thecompleteuniversityguide.co.uk/league-tables/university-and-subject-league-tables-methodology/

De Witte, K.D., López-Torres, L., 2015. Efficiency in education: A review of literature and a way forward. Journal of the Operational Research Society 68, 339–363. https://doi.org/10.1057/jors. 2015.92

Debreu, G., 1951. The coefficient of resource utilization. Econometrica 19, 273–292. https://doi.org/10.2307/1906814

Department for Education, 2018. Secondary accountability measures - Guide for maintained secondary schools, academies and free schools.

Department for Education, 2017. Teaching Excellence and Student Outcomes Framework (TEF): Lessons learned from Year Two.

Deprins, D., Simar, L., Tulkens, H., 1984. Measuring labor-efficiency in post offices, in: The performance of public enterprises: Concepts and measurement. Springer, Boston, MA, pp. 243–267. https://doi.org/10.1007/978-0-387-25534-7_16

Despotis, D.K., 2005. A reassessment of the human development index via data envelopment analysis. Journal of the Operational Research Society 56, 969–980. https://doi.org/10.1057/ palgrave.jors.2601927

Draper, D., Gittoes, M., 2004. Statistical analysis of performance indicators in UK higher education. Journal of the Royal Statistical Society. Series A (Statistics in Society) 167, 449–474.

Duerksen, C.J., 1983. Environmental regulation of industrial plant siting: How to make it work better.

Washington DC: The Conservation Foundation.

Duncombe, W., Miner, J., Ruggiero, J., 1997. Empirical evaluation of bureaucratic models of inefficiency. Public Choice 93, 1–18.

Edvardsen, D.F., Førsund, F.R., Kittelsen, S.A.C., 2017. Productivity development of Norwegian institutions of higher education 2004–2013. Journal of the Operational Research Society 68, 399–415. https://doi.org/10.1057/s41274-017-0183-x

Erdoğmuş, N., Esen, M., 2016. Classifying universities in Turkey by hierarchical cluster analysis. Education and Science 41, 363–382. https://doi.org/10.15390/EB.2016.6232

European Centre for Strategic Management of Universities, 2010. A university benchmarking handbook – Benchmarking in European higher education.

Färe, R., Grosskopf, S., 1992. Malmquist productivity indexes and Fisher ideal indexes. Economic Journal 102, 158–160.

Färe, R., Grosskopf, S., Norris, M., Zhang, Z., 1994. Productivity Growth, Technical Progress, and Efficiency Change in Industrialized Countries. The American Economic Review 84, 66–83.

Färe, R., Grosskopf, S., Weber, W.L., 1989. Measuring school district performance. Public Finance Quarterly 17, 409–428. https://doi.org/10.1177/109114218901700404

Farrell, M.J., 1959. The convexity assumption in the theory of competitive markets. Journal of Political Economy 67, 377–391.

Farrell, M.J., 1957. The measurement of productive efficiency. Journal of the Royal Statistical Society. Series A (General) 120, 253. https://doi.org/10.2307/2343100

Farsi, M., Filippini, M., Greene, W., 2005. Efficiency measurement in network industries: Application to the Swiss railway companies. J Regul Econ 28, 69–90. https://doi.org/10.1007/s11149-005-2356-9

Goldstein, H., Spiegelhalter, D.J., 1996. League tables and their limitations: Statistical issues in comparisons of institutional performance. Journal of the Royal Statistical Society. Series A (Statistics in Society) 159, 385. https://doi.org/10.2307/2983325

Greene, W., 2005. Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. Journal of Econometrics, Current developments in productivity and efficiency measurement 126, 269–303. https://doi.org/10.1016/j.jeconom.2004.05.003

Greene, W., 2002. Alternative panel data estimators for stochastic frontier models.

Greene, W., 2001. Fixed and random effects in nonlinear models.

Greene, W., 1990. A gamma-distributed stochastic frontier model. Journal of Econometrics 46, 141–163. https://doi.org/10.1016/0304-4076(90)90052-U

Grosskopf, S., Moutray, C., 2001. Evaluating performance in Chicago public high schools in the wake of decentralization. Economics of Education Review 20, 1–14. https://doi.org/10.1016/S0272-7757(99)00065-5

Guardian University Guide, 2016. Methodology behind the Guardian University Guide 2019. URL https://www.theguardian.com/education/2016/may/23/how-to-use-the-guardian-university-guide-2017

Haelermans, C., De Witte, K., 2012. The role of innovations in secondary school performance – Evidence from a conditional efficiency model. European Journal of Operational Research 223, 541–549. https://doi.org/10.1016/j.ejor.2012.06.030

Halkos, G.E., Tzeremes, N.G., 2010. Measuring regional economic efficiency: the case of Greek prefectures. The Annals of Regional Science 45 (3), 603-632

Halkos, G.E., Tzeremes, N.G., Kourtzidis, S.A., 2014. A unified classification of two-stage DEA

models. Surveys in Operations Research and Management Science 19, 1–16. https://doi.org/10.1016/j.sorms.2013.10.001

Halme, M., Joro, T., Korhonen, P., Salo, S., Wallenius, J., 1999. A value efficiency approach to incorporating preference information in data envelopment analysis. Management Science 45, 103–115.

Hanushek, E., 1971. Teacher characteristics and gains in student achievement: Estimation using micro data. The American Economic Review 61, 280–288.

Hanushek, E., 1992. The trade-off between child quantity and quality. Journal of Political Economy 100, 84–117.

Harvey, L., 2004. Analytic Quality Glossary [WWW Document]. URL http://www.qualityresearch international.com/glossary/programmeaccreditation.htm

Hashimoto, A., Ishikawa, H., 1993. Using DEA to evaluate the state of society as measured by multiple social indicators. Socio-Economic Planning Sciences 27, 257–268. https://doi.org/10.1016/0038-0121(93)90019-F

Hazelkorn, E., 2011. Rankings and the reshaping of higher education - The battle for world-class excellence. Palgrave Macmillan.

Her Majesty's Prison and Probation Service, 2017. Prison annual performance ratings 2016/2017, July.

Her Majesty's Prison and Probation Service, 2017c. Prison rating system dataset 2016 to 2017. July.

Higher Education Statistics Agency, 2010. Benchmarking to improve efficiency – Status Report November 2010.

Hildreth, C., Houck, J.P., 1968. Some estimators for a linear model with random coefficients. Journal of the American Statistical Association 63, 584–595. https://doi.org/10.2307/2284029

Hjalmarsson, L., Kumbhakar, S.C., Heshmati, A., 1996. DEA, DFA and SFA: A comparison. Journal of Productivity Analysis 7, 303–327.

Horrace, W.C., Parmeter, C.F., 2018. A Laplace stochastic frontier model. Econometric Reviews 37, 260–280. https://doi.org/10.1080/07474938.2015.1059715

Jatmiko, W., Suroso, A., Suprayitno, G., 2017. Grouping of universities in Indonesia based on data base of higher education. International Journal of Scientific and Research Publications 7.

Johnes, G., 2013. Efficiency in English higher education institutions revisited: a network approach. Economics Bulletin 33, 2698–2706.

Johnes, G., Johnes, J., 2004. International handbook on the economics of education. Edward Elgar Publishing.

Johnes, G., Schwarzenberger, A., 2011. Differences in cost structure and the evaluation of efficiency: the case of German universities. Education Economics 19, 487–499. https://doi.org/10.1080/09645291003726442

Johnes, J., 2008. Efficiency and productivity change in the English higher education sector from 1996/97 to 2004/5. The Manchester School 76, 653–674. https://doi.org/10.1111/j.1467-9957.2008.01087.x

Johnes, G., Johnes, J., 2009. Higher education institutions' costs and efficiency: Taking the decomposition a further step. Economics of Education Review 28, 107–113. https://doi.org/10.1016/j.econedurev.2008.02.001

Johnes, J., Portela, M., Thanassoulis, E., 2017. Efficiency in education. Journal of the Operational Research Society 68, 331–338. https://doi.org/10.1057/s41274-016-0109-z

Jondow, J., Lovell, C., Materov, I. & Schmidt, P. 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. Journal of Econometrics, 19, 233-238.

Joumady, O., Ris, C., 2005. Performance in European higher education: A non-parametric production frontier approach. Education Economics 13, 189–205. https://doi.org/10.1080/09645290500031215

Kerstens, K., Borger, B., Moesen, W., Vanneste, J., 1994. A non-parametric Free Disposal Hull (FDH) approach to technical efficiency: An illustration of radial and graph efficiency measures and some sensitivity results. Swiss Journal of Economics and Statistics (SJES) 130, 647–667.

Kim, H.H., Lalancette, D., 2013. Literature review on the value-added measurement in higher education. OECD Paris.

Kirjavainen, T., Loikkanen, H.A., 1998. Efficiency differences of Finnish senior secondary schools: An application of DEA and Tobit analysis. Economics of Education Review 17, 377–394. https://doi.org/10.1016/S0272-7757(97)00048-4

Kumbhakar, S., Lovell, C.A.K., 2000. Stochastic frontier analysis. Stochastic Frontier Analysis. https://doi.org/10.1017/CBO9781139174411

List, J.A., McHone, W.W., 2000. Ranking state environmental outputs: Evidence from panel data. Growth and Change 31, 23–39. https://doi.org/10.1111/0017-4815.00117

Liu, O.L., 2011a. Measuring value-added in higher education: Conditions and caveats - Results from using the Measure of Academic Proficiency and Progress (MAPP). Assessment & Evaluation in Higher Education 36, 81–94.

Liu, O.L., 2011b. Value-added assessment in higher education: A comparison of two methods. High Educ 61, 445–461. https://doi.org/10.1007/s10734-010-9340-8

Lovell, C.A.K., Walters, L.C., Wood, L.L., 1994. Stratified models of education production using modified DEA and regression analysis, in: Data envelopment analysis: Theory, methodology, and applications. Springer, Dordrecht, pp. 329–351. https://doi.org/10.1007/978-94-011-0637-5_17

Malmquist, S., 1953. Index numbers and indifference surfaces. Trabajos de Estadistica 4, 209–242. https://doi.org/10.1007/BF03006863

McCarty, T., Yaisawarng, S., 1993. Technical efficiency in New Jersey school districts, in: The measurement of productive efficiency: Techniques and applications. Oxford: Open University Press, pp. 271–287.

Meade, P., 2007. A guide to benchmarking. University of Otago.

Meeusen, W., van den Broeck, J., 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. International Economic Review 18, 435. https://doi.org/10.2307/2525757

Meredith, M., 2004. Why do universities compete in the ratings game? An empirical analysis of the effects of the U.S. News and World Report college rankings. Research in Higher Education 45, 443–461. https://doi.org/10.1023/B:RIHE.0000032324.46716.f4

Monks, J., Ehrenberg, R.G., 1999. The impact of U.S. News & World Report college rankings on admissions outcomes and pricing policies at selective private institutions. Cambridge, MA: National Bureau of Economic Research, NBER Working Paper Series No. 722 17.

Murias, P., de Miguel, J.C., Rodríguez, D., 2008. A composite indicator for university quality assessment: The case of Spanish higher education system. Social Indicators Research 89, 129–146. https://doi.org/10.1007/s11205-007-9226-z

Murias, P., Martinez, F., Miguel, C.D., 2006. An economic wellbeing index for the Spanish provinces: A data envelopment analysis approach. Soc Indic Res 77, 395–417. https://doi.org/10.1007/s11205-005-2613-4

Murnane, R.J., 1975. The impact of school resources on the learning of inner city children. Cambridge, Mass., Ballinger Pub. Co.

Mutz, R., Bornmann, L., Daniel, H.-D., 2017. Are there any frontiers of research performance? Efficiency measurement of funded research projects with the Bayesian stochastic frontier analysis for count data. Journal of Informetrics 11, 613–628. https://doi.org/10.1016/j.joi.2017.04.009

Oliveira, M.A., Santos, C., 2005. Assessing school efficiency in Portugal using FDH and bootstrapping. Applied Economics 37, 957–968. https://doi.org/10.1080/00036840500061095

Orea, L., Kumbhakar, S.C., 2004. Efficiency measurement using a latent class stochastic frontier model. Empirical Economics 29, 169–183. https://doi.org/10.1007/s00181-003-0184-2

Ofsted, 2015. The common inspection framework: education, skills and early years.

Parmeter, C., Kumbhakar, S., 2014. Efficiency analysis: A primer on recent advances. Foundations and Trends in Econometrics 7, 191–385. https://doi.org/10.1561/0800000023

Parteka, A., Wolszczak-Derlacz, J., 2013. Dynamics of productivity in higher education: cross-European evidence based on bootstrapped Malmquist indices. J Prod Anal 40, 67–82. https://doi.org/10.1007/s11123-012-0320-0

Pereira, M.C., Moreira, S., 2007. A stochastic frontier analysis of secondary education output in Portugal (No. w200706), Working Papers. Banco de Portugal, Economics and Research Department.

Pérez-Peña, R., Slotnik, D.E., 2012. Gaming the college rankings. The New York Times. URL https://www.nytimes.com/2012/02/01/education/gaming-the-college-rankings.html

Pinsent, A., Blake, I.M., Basáñez, M.G., Gambhir, M., 2016. Mathematical modelling of trachoma transmission, control and elimination, in: Advances in parasitology, Mathematical models for neglected tropical diseases. Academic Press, pp. 1–48. https://doi.org/10.1016/bs.apar.2016.06.002

Podinovski, V.V., 2004. Bridging the gap between the constant and variable returns-to-scale models: selective proportionality in data envelopment analysis. Journal of the Operational Research Society 55, 265–276. https://doi.org/10.1057/palgrave.jors.2601691

Ray, S.C., 1991. Resource-use efficiency in public schools: A study of Connecticut data. Management Science 37, 1620–1628.

Ritter, C., Simar, L., 1997. Pitfalls of normal-gamma stochastic frontier models. Journal of Productivity Analysis 8, 167–182.

Ruggiero, J., 1999. Efficiency estimation and error decomposition in the stochastic frontier model: A Monte Carlo analysis. European Journal of Operational Research 115, 555–563. https://doi.org/10.1016/S0377-2217(98)00245-8

Sarrico, C.S., Hogan, S.M., Dyson, R.G., Athanassopoulos, A.D., 1997. Data envelopment analysis and university selection. The Journal of the Operational Research Society 48, 1163. https://doi.org/10.2307/3010747

Schnedler, W., 2005. Likelihood estimation for censored random vectors. Econometric Reviews 24, 195–217. https://doi.org/10.1081/ETC-200067925

Schwartz, A.E., Zabel, J.E., 2005. The good, the bad and the ugly: Measuring school efficiency using school production functions, in: Measuring school performance & efficiency. Social Science Research Network, Rochester, NY, pp. 37–66.

Scippacercola, S., D'Ambra, L., 2014. Estimating the relative efficiency of secondary schools by stochastic frontier analysis. Procedia Economics and Finance 17, 79–88. https://doi.org/10.1016/S2212-5671(14)00881-8

Simar, L., 2000. A general methodology for bootstrapping in non-parametric frontier models. Journal

of Applied Statistics 27. https://doi.org/10.1080/02664760050081951

Simar, L., Wilson, P.W., 2011. Two-stage DEA: Caveat emptor. J Prod Anal 36, 205. https://doi.org/10.1007/s11123-011-0230-6

Simar, L., Wilson, P.W., 2007. Estimation and inference in two-stage, semi-parametric models of production processes. Journal of Econometrics 136, 31–64. https://doi.org/10.1016/j.jeconom.2005.07.009

Simar, L., Wilson, P.W., 2000. Statistical inference in nonparametric frontier models: The state of the art. Journal of Productivity Analysis 13, 49–78. https://doi.org/10.1023/A:1007864806704

Simar, L., Wilson, P.W., 1998. Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. Management Science 44, 49–61.

Steedle, J., 2012. Selecting value-added models for postsecondary institutional assessment. Assessment & Evaluation in Higher Education 37, 637–652. https://doi.org/10.1080/02602938.2011.560720

Stevenson, R.E., 1980. Likelihood functions for generalized stochastic frontier estimation. Journal of Econometrics 13, 57–66.

Storrie, D., Bjurek, H, 2000. Benchmarking European labour market performance with efficiency frontier techniques. Discussion papers, Research Unit: Labor Market Policy and Employment FS I 00-211, WZB Berlin Social Science Centre.

Thrall, R.M., 2000. Measures in DEA with an application to the Malmquist Index. Journal of Productivity Analysis 13, 125–137. https://doi.org/10.1023/A:1007800830737

Times Higher Education Rankings, 2017. World University Rankings 2018 methodology, URL https://www.timeshighereducation.com/world-university-rankings/methodology-world-university-rankings-2018

Titus, M.A., Eagan, K., 2016. Examining production efficiency in higher education: The utility of stochastic frontier analysis, in: Higher education: Handbook of theory and research. Springer, Cham, pp. 441–512. https://doi.org/10.1007/978-3-319-26829-3_9

Tone, K., Tsutsui, M., 2009. Network DEA: A slacks-based measure approach. European Journal of Operational Research, 2009, vol. 197, issue 1, 243-252

Tsionas, E., 2002. Stochastic frontier models with random coefficients. Journal of Applied Econometrics 17, 127–147. https://doi.org/10.1002/jae.637

Turner, D., 2005. Benchmarking in universities: league tables revisited. Oxford Review of Education 31, 353–371. https://doi.org/10.1080/03054980500221975

van den Broeck J., Koop G., Osiewalski J., Steel M. F. J. Stochastic frontier models: a Bayesian perspective. Journal of Econometrics. 1994;61(2):273–303. doi: 10.1016/0304-4076(94)90087-6.

Vlăsceanu, L., Grunberg, L., Pârlea, D., 2004. Quality assurance and accreditation: A glossary of basic terms and definitions. http://lst-iiep.iiep-unesco.org/cgi-bin/wwwi32.exe/[in=epidoc1.in]/?t2000=024133/(100).

Von Haldenwang, C., Ivanyna, M., 2015. Assessing the tax performance of developing countries.

Wolszczak-Derlacz, J., 2017. An evaluation and explanation of (in)efficiency in higher education institutions in Europe and the U.S. with the application of two-stage semi-parametric DEA. Research Policy 46, 1595–1605. https://doi.org/10.1016/j.respol.2017.07.010

Worthington, A.C., 2001. An empirical survey of frontier efficiency measurement techniques in education. Education Economics 9, 245–268.

Yatchew, A., 1998. Nonparametric regression techniques in economics. Journal of Economic Literature 36, 669–721.

Zhu, J., 2001. Multidimensional quality-of-life measure with an application to Fortune's best cities. Socio-Economic Planning Sciences 35, 263–284. https://doi.org/10.1016/S0038-0121(01)00009-X

Appendix A - Technical exposition

Box A1. DEA

The measure of the efficiency of the unit under examination, say unit o, relative to a set of peer units proposed in Charnes *et al.* (1978) is obtained as the maximum of the ratio of weighted outputs to weighted inputs, subject to the condition that the similar ratios for every DMU be less than or equal to unity. Furthermore, a non-negative constraint must hold for the weights. Formally (input-oriented ratio form of DEA model),

$$\max_{u,v} = \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}}$$

subject to

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1, j = 1, \dots, n$$

and

$$u_r, v_i \geq 0, \forall r, i$$

Where

n is the number of DMUs

s is the number of outputs

m is the number of inputs

y_{rj} is the amount of output *r* from unit *j*

u_r is the weight given to output *r*

x_{ij} is the amount of input i from unit *j*

v_i is the weight given to input *i*

Inputs x_{ij} and outputs y_{rj} are constants, usually observations from past decisions on inputs and the outputs that resulted from them. Input and output weights, v_i and u_r respectively, are determined objectively by the solution of this problem. Also, they are determined by the data on all of the DMUs which are being used as a reference set. Unit o is said to be efficient

$$\frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}} = 1$$

if no other unit or combination of units can produce more than unit o on at least one output without producing less in some other output or requiring more of at least one input.

The model above is an extended nonlinear programme and – to avoid existence of an infinite number of solutions – it can be transformed into an ordinary linear programme. This can be done by scaling the denominator of the objective function equal to a constant, such as 1:

(input-oriented multiplier form of DEA model)

$$\max_{\mu,v} = \sum_{r=1}^{s} \mu_r y_{ro}$$

subject to

$$\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0, j = 1, \dots, n$$

$$\sum_{i=1}^{m} v_i x_{io} = 1$$

and

$$\mu_r, v_i \geq 0, \forall r, i$$

From Charnes and Cooper (1962), we learn that the corresponding dual linear programme of the above is:

(input-oriented envelopment form of DEA model)

$$\min_{\theta_o, \lambda} \theta_o$$

subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta_o x_{io}, \qquad i = 1,2, \dots, m \quad (1$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro}, \qquad r = 1,2, \dots, s \qquad (2)$$

$$\lambda_j \geq 0, \qquad j = 1,2, \dots, n$$

This programme is called input-oriented envelopment form of the DEA model and determines whether it is possible to find a set of weights so that there exists a convex combination of the units in the sample that performs better (i.e. more outputs) than the unit under examination. In other words, the dual is seeking the efficiency rating, minimise θ_o, subject to the constraint (1) that the weighted sum of the inputs of the other DMUs is less than or equal to the inputs of the DMU being evaluated and (2) that the weighted sum of the outputs of the other DMUs is greater than or equal to the DMU being evaluate, while the weights are the values in λ_j. Notice that $\theta_o \leq 1$ and a larger value of θ_o correspond to higher efficiency. Thus, unit o is efficient when $\theta_o = 1$, but if $\theta_o < 1$ the same output levels can be achieved by using less inputs.

The output-oriented ratio, multiplier and envelopment forms of the DEA model can be obtained by following similar steps. We start from the model below, which is the reciprocal (inefficiency) measure of the input-oriented version:

(output-oriented ratio form of DEA model)

$$\min_{u,v} \frac{\sum_{i=1}^{m} v_i x_{io}}{\sum_{r=1}^{s} u_r y_{ro}}$$

subject to

$$\frac{\sum_{i=1}^{m} v_i x_{ij}}{\sum_{r=1}^{s} u_r y_{rj}} \geq 1, j = 1, \dots, n$$

and

$$u_r, v_i \geq 0, \forall r, i$$

Setting the denominator of the objective function equal to 1 yields:

(output-oriented multiplier form of DEA model)

$$\min_{\mu,v} \sum_{i=1}^{m} v_i x_{io}$$

subject to

$$\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} \geq 0, j = 1, \dots, n$$

$$\sum_{r=1}^{s} u_r y_{ro} = 1$$

and

$u_r, v_i \geq 0, \forall r, i.$

Also, the corresponding dual linear program of the above is:

(output-oriented envelopment form of DEA model)

$$\max_{\varphi_o, \lambda} \varphi_o$$

subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{io}, \qquad i = 1,2, \dots, m$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq \varphi_o y_{ro}, \qquad r = 1,2, \dots, s$$

$$\lambda_j \geq 0, \qquad j = 1,2, \dots, n$$

The output-oriented DEA model in envelopment form determines whether it is possible to find a set of weights so that there exists a convex combination of the units in the sample that performs better (i.e. less inputs) than unit o. Put differently, this dual seeks the efficiency rating, maximise φ_o, subject to the constraints (1) and (2). Notice that in this case $\varphi_o \leq 1$ and a lower value of φ_o correspond to higher efficiency. Hence, unit o is efficient when $\varphi_o = 1$, but if $\varphi_o > 1$ the same inputs can be used to achieve more output levels.

Box A2. Constant and variable returns to scale

A graphic representation of the standard DEA for the case of one input and one output under both CRS and VRS is illustrated in *Figure 3*. The frontier defines the full capacity output given the level of fixed inputs. With constant returns to scale, the frontier is defined only by unit A, with all other DMUs lying below the frontier. With variable returns to scale, the frontier is defined by units A, C and D, and only unit B falls below the frontier. Notice that the capacity output corresponding to variable returns-to-scale is lower than the capacity output corresponding to constant returns to scale.

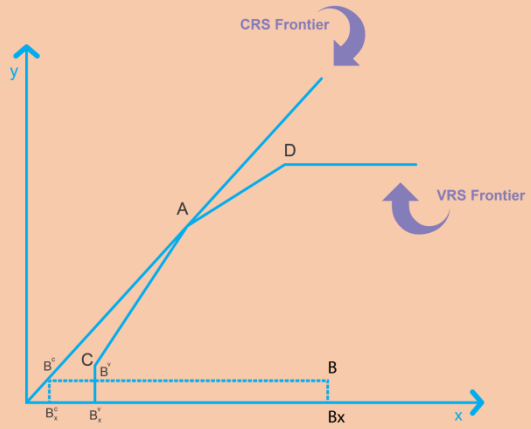


Figure 3: CRS and VRS frontiers

Adapted from Aleskerov et al. (2017)

The additional constraint mentioned previously requires that the values of the weights λ_j, $j = 1,2,\dots,n$, add up to one (Banker *et al.*, 1984). Then, the latter envelopment programmes become:

DEA VRS models

input-oriented envelopment form

$$\min_{\theta_o,\lambda} \theta_o$$

subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta_o x_{io}, \qquad i = 1,2,\dots,m$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro}, \qquad r = 1,2,\dots,s$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

and

$$\lambda_j \geq 0, \qquad j = 1,2,\dots,n$$

output-oriented envelopment form

$$\max_{\varphi_o,\lambda} \varphi_o$$

subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{io}, \qquad i = 1,2,\dots,m$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq \varphi_o y_{ro}, \qquad r = 1,2,\dots,s$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

and

$$\lambda_j \geq 0, \qquad j = 1,2,\dots,n$$

Box A3. Scale efficiency

DEA enables the researcher to obtain the scale efficiency of B by dividing the efficiency calculated under CRS to the efficiency calculated under VRS. According to *Figure 3*, the technical efficiency of unit B under CRS and VRS is equal to $\frac{OB_x^C}{OB_x}$ and $\frac{OB_x^V}{OB_x}$ respectively. Then, the scale efficiency of B can be written as

$$SE = \frac{CRS_{TE}}{VRS_{TE}} = \frac{\frac{OB_x^C}{OB_x}}{\frac{OB_x^V}{OB_x}} = \frac{OB_x^C}{OB_x^V}$$

or

$$CRS_{TE} = VRS_{TE} \times SE.$$

The latter implies that technical efficiency under the CRS condition can be decomposed in two parts: The 'pure' technical efficiency part – captured by technical efficiency under the VRS condition – and scale efficiency.

Box A4. Additive or slack-based DEA models

Unit o is said to have input slack s_i^- when input x_i can be decreased without changing the outputs and output slack s_r^+ when output y_r can be increased without changing the inputs. Then, o is said to be DEA efficient if 1. it is on the frontier (that is, $\theta_o = 1$ or $\varphi_o = 1$) and 2. its slacks s_i^-, s_r^+ are zero. On the contrary, when o is on the frontier and some of its slacks are non-zero, it called DEA weakly efficient.

The latter suggests that slack calculation is a two-stage process, because it requires obtaining the efficiency of the unit under examination first. That is, the researcher calculates θ_o or φ_o by ignoring the slacks (i.e. solve the envelopment form of the DEA models) and then optimises the slacks by keeping θ_o or φ_o in the following programming problems fixed:

DEA CRS models with slacks

input-oriented envelopment form

$$\max_{s_i^-, s_r^+} \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+$$

subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \theta_o x_{io}, \qquad i = 1,2,\dots,m$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{ro}, \qquad r = 1,2,\dots,s$$

$$\lambda_j \geq 0, \qquad j = 1,2,\dots,n$$

$$s_i^- \geq 0, \qquad i = 1,2,\dots,m$$

$$s_r^+ \geq 0, \qquad r = 1,2,\dots,s$$

output-oriented envelopment form

$$\max_{s_i^-, s_r^+} \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+$$

subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io}, \qquad i = 1,2,\dots,m$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = \varphi_o y_{ro}, \qquad r = 1,2,\dots,s$$

$$\lambda_j \geq 0, \qquad j = 1,2,\dots,n$$

$$s_i^- \geq 0, \qquad i = 1,2,\dots,m$$

$$s_r^+ \geq 0, \qquad r = 1,2,\dots,s$$

DEA VRS models with slacks

input-oriented envelopment form

$$\max_{s_i^-, s_r^+} \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+$$

subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \theta_o x_{io}, \qquad i = 1,2,\dots,m$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{ro}, \qquad r = 1,2,\dots,s$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j \geq 0, \qquad j = 1,2,\dots,n$$

$$s_i^- \geq 0, \qquad i = 1,2,\dots,m$$

$$s_r^+ \geq 0, \qquad r = 1,2,\dots,s$$

output-oriented envelopment form

$$\max_{s_i^-, s_r^+} \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+$$

subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io}, \qquad i = 1,2,\dots,m$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = \varphi_o y_{ro}, \qquad r = 1,2,\dots,s$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j \geq 0, \qquad j = 1,2,\dots,n$$

$$s_i^- \geq 0, \qquad i = 1,2,\dots,m$$

$$s_r^+ \geq 0, \qquad r = 1,2,\dots,s$$

Box A5. Composite Indicators with DEA

The construction composite indicators with DEA can be directly seen by recalling the output-oriented DEA model presented above, in its multiplier form:[52]

$$\min_{\mu,v} \sum_{i=1}^{m} v_i x_{io}$$

subject to

$$\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} \geq 0, j = 1,\dots,n$$

$$\sum_{r=1}^{s} u_r y_{ro} = 1$$

$$u_r, v_i \geq 0, \forall r, i.$$

Notice that in such case $\sum_{i=1}^{m} v_i x_{io}$ consists a compilation of the partial indicators $v_1 x_{1o}, v_2 x_{2o}, \dots, v_m x_{mo}$ and can thus be perceived as a composite indicator for the unit o. Then, we can obtain the desired composite indicators by solving the programme above for every unit

[52] When it is assumed that the partial indicators should be as high as possible, the output-oriented multiplier DEA model should be used. An input-oriented multiplier DEA model is more suitable if the converse is true, that is, if all of the indicators were preferred to be as low as possible.

examined.

Box A6. Malmquist productivity index

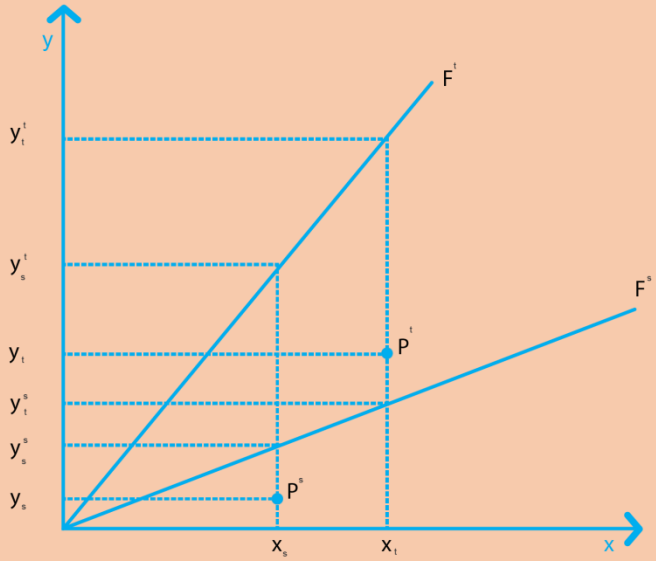


Figure 4: Production frontier change
Adapted from Johnes (2008)

According to *Figure 4, F^s* represents the CRS production frontier corresponding to the efficient levels of output y that can be produced in time s from a given level of input x. Similarly, F^t represents the CRS production frontier corresponding to the efficient levels of output *y* that can be produced in time t from a given level of input *x*. In time periods s and t, unit A produces at points $A^s$and A^t respectively. Thus, its technical efficiency corresponds to $\frac{0y_s}{0y_s^s}$ in *s* and $\frac{0y_t}{0y_t^t}$ in t. This measure is in fact the output distance function which, for the general case, can be denoted by $D_O^s(x_s, y_s)$ and $D_O^t(x_t, y_t)$ for s and t respectively (Coelli *et al.,* 2005). The time dimension allows an analysis of productivity change which can be measured by using the Malmquist productivity index. That is, on the assumption that the unit seeks to maximise output for a given level of input (i.e. an output-oriented approach), productivity change might be evaluated by comparing the efficiencies of observations A^s and A^t calculated relative to the frontier in period *s*, i.e., or by comparing the efficiencies of A^s and A^t calculated relative to the frontier in period *t,* i.e. $\frac{0y_t}{0y_t^s} \Big/ \frac{0y_s}{0y_s^s}$ and $\frac{0y_t}{0y_t^t} \Big/ \frac{0y_s}{0y_s^t}$.[53] In practice however, the Malmquist approach takes a geometric mean of these two measures. The general formula for the Malmquist productivity change index is:

$$M_O(x_t, y_t, x_s, y_s) = \left[\left(\frac{D_O^s(x_t, y_t)}{D_O^s(x_s, y_s)}\right)\left(\frac{D_O^t(x_t, y_t)}{D_O^t(x_s, y_s)}\right)\right]^{1/2}$$

[53] The evaluation of productivity change for the case of an input-oriented framework can be obtained in a similar manner. See Johnes and Johnes (2004) for more details.

where $D_O^s(x_t, y_t)$ denotes the distance of the period t observation from the period *s* frontier. When $M_O(x_t, y_t, x_s, y_s) > 1$, then productivity between the two time periods has improved. But, if $M_O(x_t, y_t, x_s, y_s) < 1$, productivity between the two time periods has *deteriorated*.

Further, as it can be seen from *Figure 4*, the change in the production position of unit *A* over the time periods *s* and t has the following underlying determinants:

- unit *A* can produce more because the sector's production frontier has moved outwards, and therefore the potential for production is expanded
- the unit's position relative to the time-relevant frontier can change.

The Malmquist productivity index, as written above, conceals these two effects. However, it is possible to decompose it into the two components as follows (Färe *et al.,* 1989; 1994):

$$M_O(x_t, y_t, x_s, y_s) = \frac{D_O^t(x_t, y_t)}{D_O^s(x_s, y_s)} \left[\left(\frac{D_O^s(x_t, y_t)}{D_O^t(x_t, y_t)} \right) \left(\frac{D_O^s(x_s, y_s)}{D_O^t(x_s, y_s)} \right) \right]^{1/2}.$$

The first component

$$EC = \frac{D_O^t(x_t, y_t)}{D_O^s(x_s, y_s)}$$

is technical efficiency and reflects changes in the relative efficiency of the unit A (e.g. DMUs getting closer to or further away from the efficiency frontier), while the second component is technological change

$$TC = \left[\left(\frac{D_O^s(x_t, y_t)}{D_O^t(x_t, y_t)} \right) \left(\frac{D_O^s(x_s, y_s)}{D_O^t(x_s, y_s)} \right) \right]^{1/2},$$

measures the shift in the production frontier itself and reflects effects that concern the sector as a whole. Values greater than one for both EC and TC suggest improvement, while values of less than one suggest the opposite.

Box A7. Free disposal hull

The assumption that FDH relaxes is the convexity assumption. It is desirable to relax the convexity assumption as it is frequently difficult to find a good empirical or theoretical justification for convex production sets in efficiency analysis. Farrell (1959) indicates that indivisibility of inputs and outputs, economies of scale and specialisation are possible violations of convexity[54].

The free disposal assumption that fewer outputs can always be produced using more inputs results in a production possibility set with a stepwise shape. Such a FDH production possibility set is shown in the graph below.

[54] It is important to note that both DEA and FDH approaches yield consistent estimators when the true production set is convex.

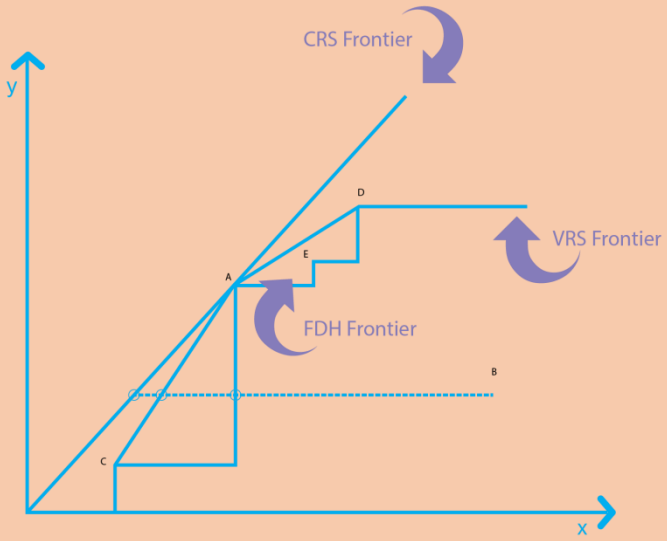


Figure 5: CRS, VRS and FDH frontiers

Notice that the FDH production possibility set consists of a subset of the production possibility set either under the constant or the variable returns-to-scale condition.

In FDH models, the final constraint on the semi-positivity of λ_j is replaced by $\lambda_j \in \{0,1\}$, suggesting that λ_j is binary. Formally,

(input-oriented FDH model)

$$\min_{\theta_o,\lambda} \theta_o$$

subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta_o x_{io}, \qquad i = 1,2,\dots,m$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro}, \qquad r = 1,2,\dots,s$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

and

$$\lambda_j \in \{0,1\}, \qquad j = 1,2,\dots,n$$

(output-oriented FDH model)

$$\max_{\varphi_o,\lambda} \varphi_o$$

subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{io}, \qquad i = 1,2,\dots,m$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq \varphi_o y_{ro}, \qquad r = 1,2,\dots,s$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

and

$$\lambda_j \in \{0,1\}, \qquad j = 1,2,\dots,n$$

Box A8. SFA

Aigner *et al.* (1977) and Meeusen and van den Broeck (1977) proposed, almost simultaneously, a formulation within which the error term is composed of two elements: one element represents natural random variation in performance (i.e. the noise) and the other element represents the (in)efficiency factor. The former element could be of either sign, while the latter is non-negative. The resulting frontier is presented in terms of a general production function, known as a stochastic production frontier. For instance, if we assume that $f(x_i;\beta)$ takes the log-linear Cobb-Douglas form, the stochastic frontier model can be written as:

(additive form)

$$\ln y_i = \beta_0 + \sum_{i=1}^{n} \beta_i ln x_{ni} + v_i - u_i$$

or

$$y_i = \exp(\beta_0 + \sum_{i=1}^{n} \beta_i ln x_{ni} + v_i - u_i)$$

or

(multiplicative form)

$$y_i = \underbrace{\exp(\beta_0 + \sum_{i=1}^{n} \beta_i ln x_{ni})}_{\text{deterministic component}} \times \underbrace{\exp(v_i)}_{\text{noise}} \times \underbrace{\exp(-u_i)}_{\text{inefficiency}}$$

where

y_i is the observed output of i unit ($i = 1,\dots,n$)

β are the unknown production technology parameters to be estimated

x_{ni} are the m inputs

v_i captures the stochastic elements and unobserved heterogeneity (noise)

u_i is the inefficiency term and captures the shortfall in output given the inputs

The model above implies that unit i is 100% efficient if $u_i = 0$, but there is some inefficiency when $u_i > 0$. Also, it follows that the total error term of the model has two components: the 'noise'

component v_i and the inefficiency component u_i. More formally,

$$\varepsilon_i = v_i - u_i.$$

For this reason, the error term ε_i in SFA is referred to as the composite (or combined) error term. It is further assumed that each v_i is distributed independently of each u_i and that both errors are uncorrelated with the explanatory variables in x_{ni}. Moreover, it is assumed that the stochastic component (noise) v_i is independent and identically distributed (*iid*) normal random variable with mean zero and constant variance σ_v^2:

$$\text{A1. } v_i \sim iid\ N(0, \sigma_v^2),\ i = 1, \dots, n\ .$$

This, along with the fact that $u_i \geq 0$, implies that ε_i could be asymmetric. In particular, if $u_i = 0$, then $\varepsilon_i = v_i$ and the error term is symmetric, suggesting that the data do not support a technical inefficiency story. But, if $u_i > 0$, then the error term is negatively skewed, suggesting that there is evidence of technical inefficiency in the data.

Hence, assuming that there is evidence of technical inefficiency in the data, the objectives of the analysis are to obtain estimates of

- the production function technology parameters β in $f(x_i; \beta)$,
- the individual efficiencies, u_i.

Although the estimated function may be of interest on its own, we are usually more interested in the resulting estimates of the individual efficiencies. One of the most common technical efficiency measures is the ratio of observed output to the corresponding stochastic frontier output:

$$TE_i = \frac{f(x_i; \beta) \times \exp(v_i) \times \exp(-u_i)}{f(x_i; \beta) \times \exp(v_i)} = \exp(-u_i) \qquad (1)$$

This technical efficiency measure takes values between zero and one. When $TE_i = 1$, unit i is fully efficient and, correspondingly, the observed output y_i reaches its maximum obtainable value. Moreover, a $TE_i < 1$ provides a measure of the shortfall of the observed output from maximum feasible output.

The estimation strategy usually deployed consists of two steps:

Step 1. The maximum likelihood method is used to estimate the all the parameters of the model.

Step 2. Conditional on the maximum likelihood estimates obtained in the previous step, the technical efficiency is estimated for each unit i by decomposing the maximum likelihood residual term ε_i into a noise component and a technical efficiency component.

As far as the second step is concerned, the extraction of separate estimates for v_i and u_i from ε_i requires, on top of A1., distributional assumptions for the technical inefficiency component u_i.

Distributional assumptions for the technical inefficiency component

Half normal distribution

We first examine the case when the technical inefficiency component is independently and identically distributed half normal random variable with mean zero and constant variance σ_u^2,

$$A2. u_i \sim iid\ N^+(0, \sigma_u^2),\ i = 1, \dots, n\ ,$$

an assumption initially adopted by Aigner *et al.* (1977). N_+ denotes a half-normal distribution, i.e. a truncated normal distribution where the point of truncation is 0 and the distribution is concentrated on the half-interval [0,+∞). A2. is based on the plausible proposition that the modal value of the technical inefficiency is zero, with increasing values of technical inefficiency becoming increasingly less likely. This assumption is also based on tractability. It is relatively easy to derive

the distribution of the sum of v_i and u_i under A1.-A2.

Aigner *et al.* (1977) parametrise the log likelihood function for this case, also called normal-half normal model, in terms of $\sigma = (\sigma_v^2 + \sigma_u^2)^{1/2}\ and\ \lambda = \frac{\sigma_u}{\sigma_v}$. Notice that the parametrisation from σ_v and σ_u to σ and λ is convenient, because λ provides a measure of the relative contributions of v_i and u_i to ε_i:

- When $\lambda \to 0$, either $\sigma_v^2 \to +\infty$ or $\sigma_u^2 \to 0$, which implies that the symmetric error component v_i dominates the one-sided error component u_i in the determination of ε_i.
- When $\lambda \to +\infty$, either $\sigma_v^2 \to 0$ or $\sigma_u^2 \to +\infty$, which implies that the one-sided error component u_i dominates the symmetric error component v_i in the determination of ε_i.

The log likelihood function for this parametrisation for a sample of N units is

$$\ln L = constant - I \ln \sigma + \sum_i^N \ln \Phi\left(-\frac{\varepsilon_i \lambda}{\sigma}\right) - \frac{1}{2\sigma^2}\sum_i^N \varepsilon_i^2,$$

where $\Phi(.)$ is the standard normal distribution function.

The next step is to obtain the estimates of the individual efficiencies. Note that the estimate of the error term ε_i does carry some information on u_i: When $\varepsilon_i > 0$, then chances are that u_i is not very large, as $E(v_i) = 0$ and $u_i \geq 0$, implying that unit i is relatively efficient. But, $\varepsilon_i < 0$ suggests that u_i will tend to be large, implying that unit i is relatively inefficient.

Hence, the prediction of technical efficiency can be based on the conditional expectation $E(u_i|\varepsilon_i)$, which contains whatever information ε_i contains concerning u_i. Jondow *et al.* (1982) showed that under A2., the conditional distribution of u given ε is:

$$f(u|\varepsilon) = \frac{f(u,\varepsilon)}{f(\varepsilon)} = \frac{1}{\sqrt{2\pi}\sigma_*} \frac{\exp(-\frac{u-\mu_*}{2\sigma_*^2})^2}{1 - \Phi(-\frac{\mu_*}{\sigma_*})}$$

where

$$\mu_* = \frac{-\varepsilon\sigma_u^2}{\sigma^2}\ and\ \sigma_*^2 = \frac{\sigma_u^2\sigma_v^2}{\sigma^2}.$$

Further, since $f(u|\varepsilon)$ is distributed as $N^+(\mu_*, \sigma_*^2)$, its mean can serve as a point estimator of u_i[55]

$$E(u_i|\varepsilon_i) = \sigma_* \left[\frac{\varphi\left(\frac{\varepsilon_i\lambda}{\sigma}\right)}{1 - \Phi\left(\frac{\varepsilon_i\lambda}{\sigma}\right)} - \left(\frac{\varepsilon_i\lambda}{\sigma}\right)\right]$$

and estimates of the technical efficiency of each unit i can be obtained by replacing the above in (1). Unfortunately, in this case the estimates of technical efficiency are not consistent, as the variation associated with the distribution of $(u_i|\varepsilon_i)$ is independent of i. This appears to be the best that can be achieved with cross-sectional data.

Other distributional specifications

Johnes and Johnes (2009) point out that the technical inefficiency component u_i could follow any non-normal distribution – so that it can be distinguished from the noise component v_i, though, for the reasons of analytical convenience covered above, the half-normal distribution is a common

[55] The mode of the distribution $N^+(\mu_*, \sigma_*^2)$ may also be used as a point estimator of u_i. See Kumbhakar and Lovell (2000) for an extensive discussion.

assumption. Alternative distributional assumptions for u_i include[56]:

$$A3.\ u_i \sim iid\ N^+(\mu, \sigma_u^2), i = 1, \dots, n \text{ (truncated normal distribution)}$$

$$A4.\ u_i \sim iid\ G(\lambda, 0) \text{ (exponential distribution)}$$

$$A5.\ u_i \sim iid\ G(\lambda, m) \text{ (gamma distribution)}$$

The truncated normal distribution was initially proposed by Stevenson (1980). A3. generalises A2. by allowing the normal distribution (which is by definition truncated below zero) to have a non-zero mode μ. The exponential distribution originally used in Meeusen and van den Broeck (1977), was generalised by allowing u_i to follow a gamma distribution in Greene (1990). As both the truncated normal and the gamma distributions require two parameters to be estimated, rather than only one in the half-normal and the exponential distribution, they provide a more flexible presentation of the efficiency pattern in the data. The log likelihood functions and the estimators of the technical efficiency that correspond to the above specifications can be found in Kumbhakar and Lovell (2000).

Choice of distributional specifications

When efficiency is examined by applying a stochastic frontier model, the importance of the distributional assumptions on the noise component v_i and technical inefficiency component u_i constitutes an issue of special interest. It is commonly accepted that v_i is a two-sided normally distributed variable, even though some exceptions can be found in the literature[57]. With regards to u_i, not many papers examine differences in the estimates across distributional assumptions. A notable example is Greene (1990). Greene performs a comparison of average inefficiency levels across alternative distributional specifications for 123 U. S. electric utility providers and finds that the estimates are quite robust to distributional choice. Kumbhakar and Lovell (2000) produced rank correlation coefficients between pairs of efficiency estimates as low as 0.746 (exponential and gamma) and as high as 0.98 (half normal and truncated normal)[58].

Then, as proposed by Kumbhakar and Lovell (2000) and Parmeter and Kumbhakar (2014), it may prove wise to follow the advice of Ritter and Simar (1997) and assume a relatively simpler (half normal or exponential) to a more flexible specification (truncated normal or gamma) for the distribution of u_i.

The choice of distributional specifications may also depend on the available computational tools. For example, STATA can fit models in which the noise component v_i follows A1. and the technical inefficiency component u_i follows A2., A3. or A4.

Box A9. Two-stage SFA

There are a few technical issues regarding estimating the second stage of SFA. The first stage of the SFA approach requires u_i to be distributed identically and independently across all observations, yet the second stage assumes there exists a functional relationship between the efficiency represented by u_i in the first stage and the explanatory variables. Also, since the estimated efficiencies have a distribution bounded between 0 and 1 (double censored), the OLS method cannot be applied because it will give biased estimations. In such occasion, the Tobit model – proposed by Schnedler (2005) – is a more appropriate tool.

[56] For more on alternative distributional assumptions, see Parmeter and Kumbhakar (2014).

[57] See for example Horrace and Parmeter (2018).

[58] In Ruggiero (1999), rank correlations of stochastic frontier estimates were compared, under the assumption that inefficiency was either half normal (which was the true distribution) or exponential (a mis-specificated distribution) and found very little evidence that mis-specification impacted the rank correlations in any meaningful fashion.

Box A10. Latent class models

To relax the assumption that the production process or technology is the same for all producers, Orea and Kumbhakar (2004) proposed that a latent class structure be applied to SFA. The LC-SFA model is specified as:

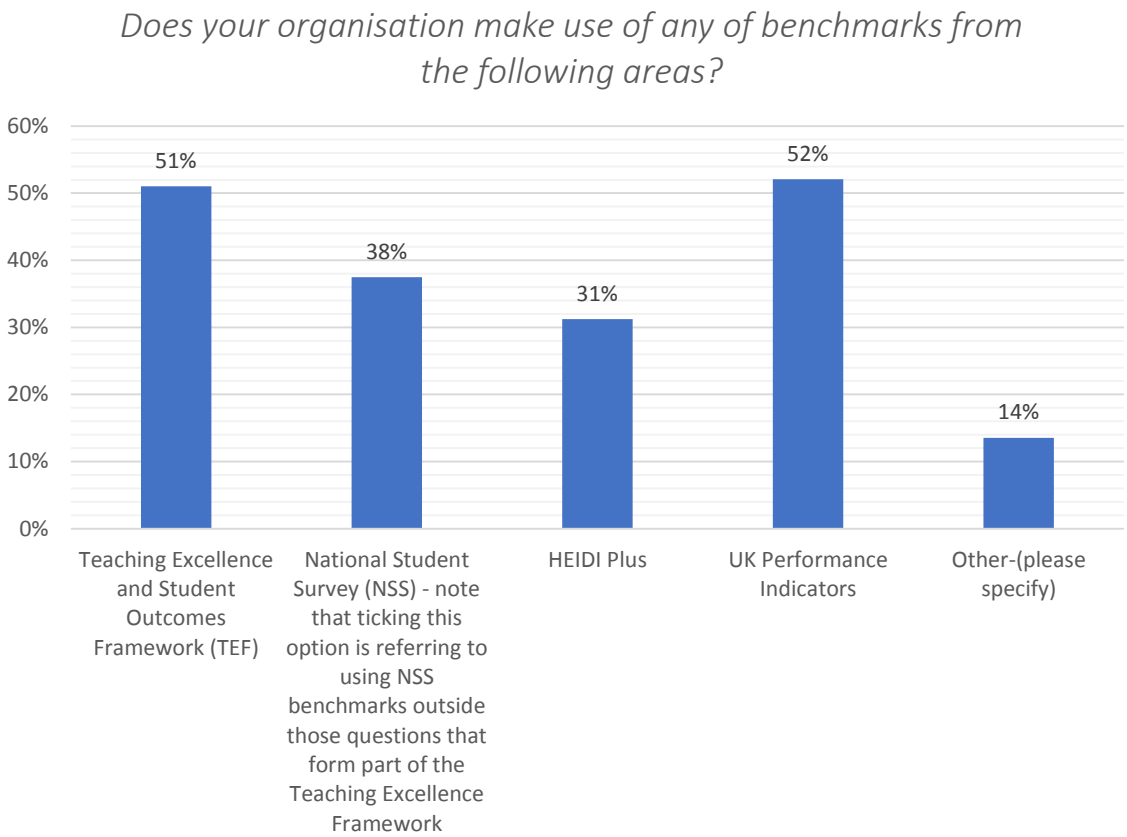
$$y_{it} = a_{it} + \beta_m x_{it} + v_{ti,m} + u_{it,m},$$

where *m* is the number of latent classes. The LC-SFA model is estimated via maximum simulated likelihood techniques.

Source: Titus and Eagan (2016)

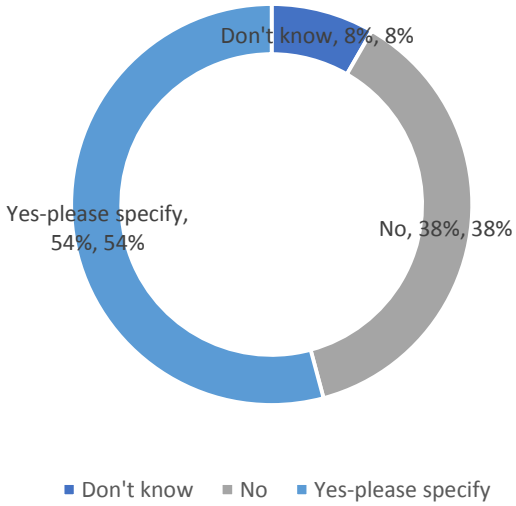
Appendix B - Survey findings

The charts and tables below present the results from the Alma Economics Benchmarking Survey, which was sent to all 138 HE planning directors via the Higher Education Strategic Planners Association (HESPA). The survey was open for three weeks and during that time one reminder was sent. We received 96 responses though 37 of these were blank so the true response rate was 43%.



The main examples reported in the 'other' category were various league tables and different data sets including UCAS, Longitudinal Educational Outcomes and the Destination of Leavers from HE (DLHE) survey.

The benchmarks listed above are calculated using the same methodology. Are there any other benchmarking methodologies your institution uses?



The different methodologies outlined mostly referred to comparisons of different sets of defined groups (e.g. various subjects, the Russell Group). Comparisons appeared to use means, top and bottom quartiles.

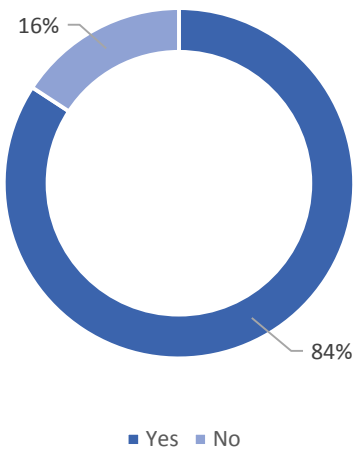
In which aspects of your institution's work are benchmarks used the most? List up to three.

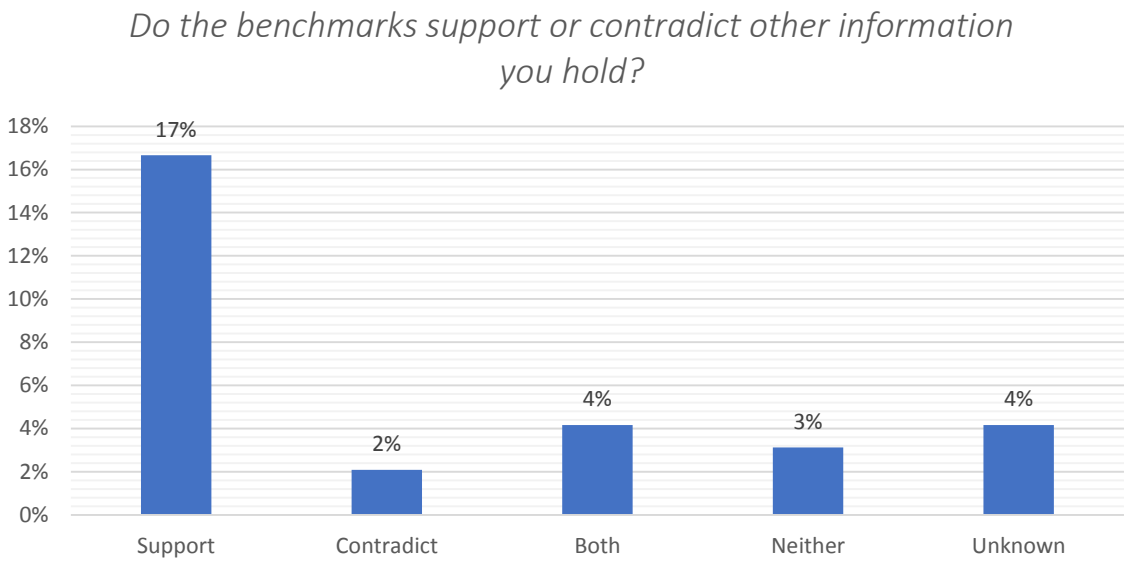
Area	No1	No2	No3
KPIs (including all performance related)	20	8	8
TEF	5	5	2
Strategic Planning	3	5	1
Widening participation monitoring	5	2	2
Access	2	2	1
Student recruitment	2	0	0
Retention	2	0	0
Progression	2	0	0
Target-setting for funders/regulators	1	2	1

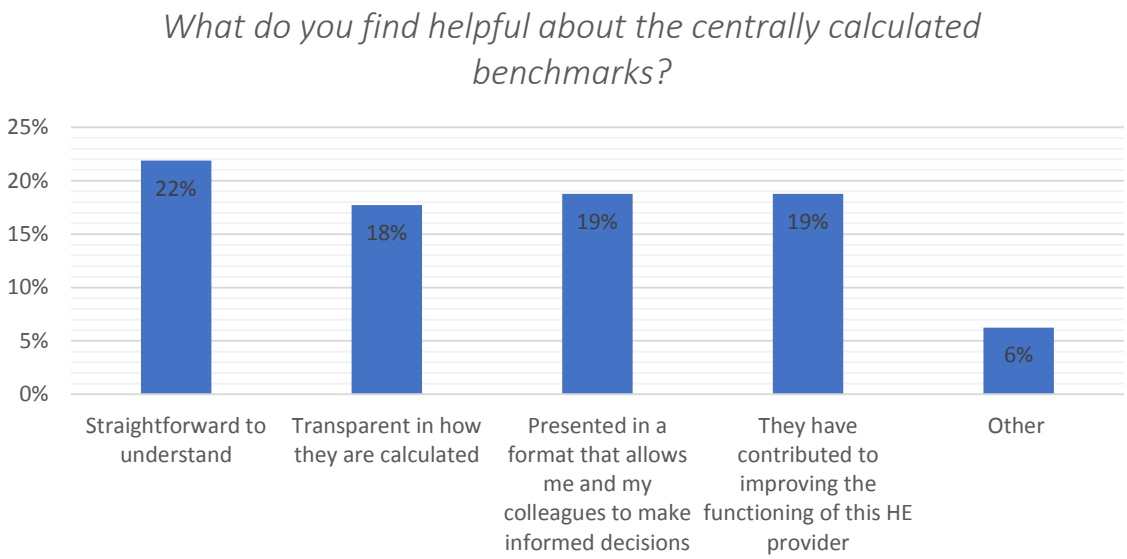
Which, if any, particular benchmarks help your institution achieve its strategic aims and why? For example, any non-continuation benchmarks within TEF, or a particular widening participation benchmark within the UK Performance Indicators.

Benchmark	Number of providers who reported making use of them
TEF	22
NSS	17
Widening participation	17
UKPI	15
Non-continuation	12
Comparisons	4
DLHE	3
Value Added	3
Employability	3

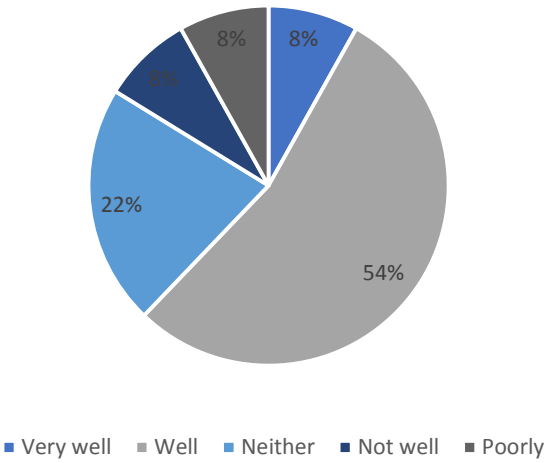
Do the centrally calculated benchmarks provide new information to your institution?

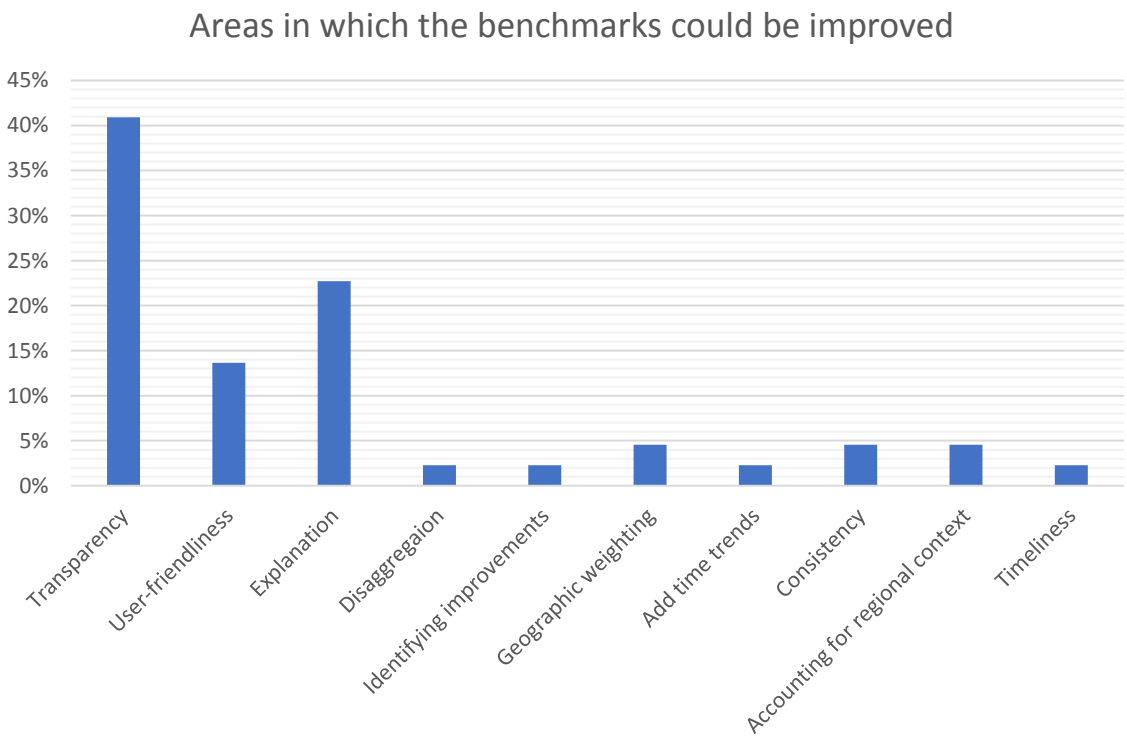


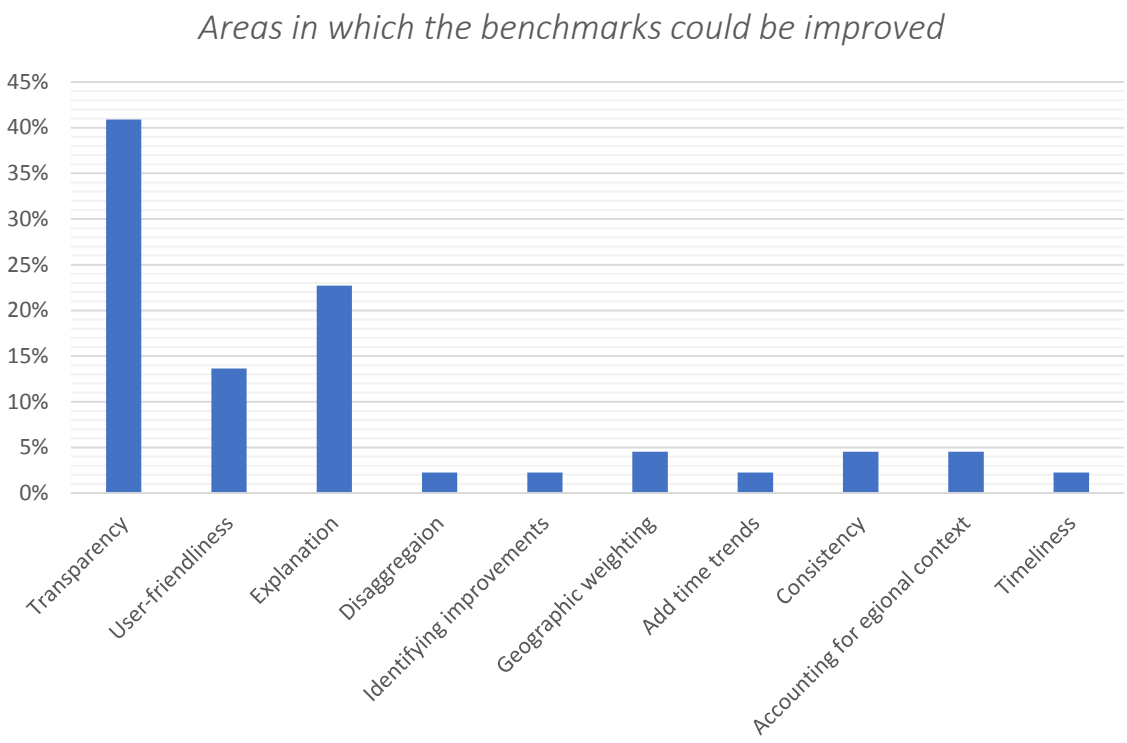

Do the benchmarks support or contradict other information you hold?
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
17%
2%
4%
3%
4%
Support
Contradict
Both
Neither
Unknown


What do you find helpful about the centrally calculated benchmarks?
25%
20%
15%
10%
5%
0%
22%
18%
19%
19%
6%
Straightforward to understand
Transparent in how they are calculated
Presented in a format that allows me and my colleagues to make informed decisions
They have contributed to improving the functioning of this HE provider
Other

How well do you feel the centrally calculated benchmarks deal with your institution's specific context?






Areas in which the benchmarks could be improved
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Transparency
User-friendliness
Explanation
Disaggregaion
Identifying improvements
Geographic weighting
Add time trends
Consistency
Accounting for egional context
Timeliness