Can we assess teaching quality on the basis of student outcomes? A Stochastic frontier application

Dimitris Giraleas

Aston Business School, Aston University, Birmingham, UK[1]

Abstract:

This paper proposes a new application of Stochastic Frontier Analysis (SFA) for estimating the student performance gap and how this can be used to assess changes of teaching quality at the individual unit-of-study level (module-level). Although there have been other examples in the literature that assess 'efficiency' in student outcomes, this is the first study that proposes the use of SFA specifically at the module level and with the goal of creating an aggregate measure of 'quality', thus avoiding the known issue of the statistical inconsistency of unit-specific SFA estimates. A case study is presented on how the approach can be applied in practice, with discussion on potential implementation issues. This paper is targeted to academics and policy makers that are interested in the quantitative assessment of student outcomes and specifically to those who want to assess how changes in module structure and/or delivery have affected said student outcomes.

Key words: Educational Production Function, Stochastic Frontier Analysis, student attainment gap, teaching quality, Higher Education

[1] Contact: Dimitris Giraleas, d.giraleas1@aston.ac.uk, Aston University, Aston Triangle, Birmingham B4 7ET, United Kingdom

1. Introduction

The importance of higher (university-level) education cannot be overstated in terms of the creation of human capital and new research that can enhance the human condition (Hanushek and Wößmann 2010). In terms of the creation of human capital, a critical factor is the overall quality of teaching provided in these institutions. Teaching quality in higher education is in fact a very topical issue in many countries (as exemplified by the Bologna Process in EU countries, the National Survey of Student Engagement (NSSE) as applied to numerous countries around the world, the Collegiate Assessment of Academic Proficiency (CAAP) and Collegiate Learning Assessment (CLA) as applied in the US); in the UK, teaching quality is in the forefront of assessing HEIs through the newly introduced Teaching Excellence Framework (TEF).

However, teaching quality is an elusive concept and there is considerable discussion on how to define teaching excellence (Gunn and Fisk 2013; Land and Gordon 2015). Brusoni et al. (2014) note that teaching excellence has multiple dimensions, some of them ‘internal’, ie specific to the individual unit of study, such as the ability of individual lectures to inspire students and communicate clearly, the material of the individual study unit and its delivery, and some of them ‘external’, ie specific to the wider institution in which the learning takes place, such as appropriate facilities and well- organised programmes of study. As such it is important to distinguish between the levels that this teaching quality characteristic refers to, from the more aggregate to the more granular: for example, teaching quality can refer to an institution as a whole, a school/department within the institution, a course/programme within a school, a single unit of study within the course/programme, or even a single lecturer/instructor within a unit of study. This paper proposes a new quantitative methodology to estimate the ‘internal’ teaching quality dimension at the unit of study level (referred to thereafter as the module level[2]), through examining the performance gap of a student cohort, in order to provide a robust, evidence-based method for assessing teaching quality and to help address whether changes in the module delivery and design have the desired impact of improving student learning.

Teaching quality is a complex subject and as such there are a number of issues that need to be addressed even before the analysis can begin. A first issue is the level of analysis: should it be institution-wide (aggregate) or more granular? Institution or school/department-wide analyses are more appropriate for accreditation or league-table purposes, while more granular analyses, ie those that focus on the module level, are more appropriate for either formative or summative assessment of teachers/lecturers (Hinchey 2010).

A second important issue to address is how to identify teaching excellence. Some systems rely on quality standards set by professional or other accreditation bodies; one of the main goals of the aforementioned Bologna Process is to define such standards and this is also the main role of the Quality Assurance Agency in Higher Education (QAA) in the UK and the Australian Learning and Teaching Council (ALTC) amongst others. Other systems rely on student surveys, such as the aforementioned NSSE. Another option is to utilise peer observations and site visits; such systems are used in Germany and it could also be argued that the Institutional Audits by the QAA in the UK also fall in this category. Lastly, teaching excellence could also be assessed through examining student outcomes; examples of such systems are the CAAP in the US and the TEF in the UK. It should be noted that teaching assessment frameworks are not necessarily confined to using a single system for identifying or describing teaching excellence. For example, the system in Germany incorporates both

[2] A module is defined as an individual unit of study that lasts a certain amount of time and has specific and clearly stated learning aims, a specified delivery method that sets out how the learning material will be delivered and a specified assessment that is constructively aligned with the learning aims. Each student is awarded a grade at the end of the module based on the assessment and the individual module grades are aggregated at the end of the programme of study to derive a student's overall degree grade.

externally-set quality standards (and internal self-assessment from the institutions) and external peer reviews. TEF in the UK relies on both student surveys (specifically the National Student Survey (NSS)) and student outcomes (sourced from the Destination of Leavers from Higher Education (DLHE) survey).

The systems available for describing teaching excellence offer a guide on how excellence could be measured. Teaching quality standards provide a baseline and teaching quality can further be revealed by peer observations, student opinion and student outcomes. All three assessment methods are important and can be used in tandem to reveal what excellence in teaching looks like, but unfortunately not all of them are practical or even possible to implement in all instances.

- Peer observations are potentially very useful in discovering and disseminating best practice but are fraught with practical and theoretical difficulties. On the practical side, an in-depth observation of a full module is a very time-consuming process, requiring both classroom observations and a detailed examination of the module structure (learning outcomes, material delivered, delivery methods). Hammersley-Fletcher and Orsmond (2004) reveal a number of the theoretical difficulties with the approach, such as whether the peer has the skills and/or the appropriate technical background necessary to provide constructive criticism. Bernstein (2008) argues that peer observations should mainly be used as formative reviews, as their reliability tends to be low.

- Student opinion, sourced through student satisfaction surveys is arguably the most heavily utilised method in assessing teaching quality in the UK (mainly through the NSS) and Australia (through the Course Experience Questionnaire (CEQ)); both surveys are taken as performance indicators at the programme/degree level, but student surveys are also used for formative or summative assessments for teachers/lectures. There is extensive research on the validity of student surveys as a tool for revealing teaching excellence. Spooren, Brockx and Mortelmans (2013) provide a comprehensive review with a focus on the link between student satisfaction, student outcomes and student bias with regards to providing feedback. According to this review, previous studies have found moderate to large positive correlations between student satisfaction scores and other indicators of teaching quality. However, the review also finds significant issues with regards to content- and construct-related validity, relating specifically to issues of self-selection bias on the part of students and correlation between satisfaction scores and teacher characteristics unrelated to effective teaching (eg teacher attractiveness).

- Student outcomes is another important metric of teaching quality; one could argue the most important, as improving teaching quality should be positively and strongly correlated with improving student outcomes. There are at least two perspectives when examining student outcomes: The first deals with outcomes that are directly observable and are considered the ‘outputs’ of the education process – examples here include module pass rates, drop-out ratios, degree outcomes (final attainment) and student employment status after the degree. The advantage of using these outputs as measures of teaching quality is that they are well-defined and unambiguous. The disadvantage however is that there are factors other than teaching quality that can significantly affect these measures, such as student characteristics (Hanushek 1979). These confounding factors have led to the proposal of alternative, process-based measures (Gibbs 2010), such as class size, student engagement and the quality of feedback to students. Such process measures however are not without issues; the

> effect of these measures on student outcomes is difficult to judge and might differ for different institutions, the definitions of some measures are open to interpretation, while other measures require the adoption of well-defined, consistent standards across all participating institutions.

The method of assessing teaching quality proposed in this study is based on student outcomes, defined as the outputs of the educational process, controlling for student characteristics and other potentially confounding factors through a multivariate regression model based on the concept of the educational production function. The analysis is applied at the module level and the unit of analysis is the individual student, since learning takes place primarily within the structured parameters of a module and students are assessed on a per module basis. The study innovates in the introduction of the concept of student attainment gap and suggests a methodology for measuring it based on econometric stochastic frontier models. The resulting metric can be used to infer changes in the quality of teaching provision over time. These issues will be discussed in detail in section 2. In short, the attainment gap is defined as the difference between the ideal learning outcome of each individual student in a given module, controlling for relevant student characteristics, and what each individual student has actually achieved. If the average attainment gap in a student cohort decreases over time, it can be inferred that module delivery (ie teaching quality) has improved, since the analysis already controls for prior student characteristics. The goal is to simply measure changes and not to explain how/why these changes affected teaching quality; this can be considered an advantage in such analyses, since there is no need to strictly define what teaching quality should look like.

The second main goal of this study is to demonstrate through a clear and comprehensive case study how this new approach can be employed in practice and provide some guidance on how the approach can be modified to fit a variety of different modules. The case study will also demonstrate how the educational production function can be used at the module level (which very few studies have attempted before, as will be discussed below) and the insights that can be gained from the analysis. The case study will also provide some quantitative evidence on the impact of student characteristics to student outcomes that will be of interest to teachers/lecturers that teach similar subjects (namely teaching quantitative subjects to business school students) in the UK; however, since this is a relatively narrow scope, this discussion is of secondary importance. The case study will be presented in section 3 of this report. Section 4 will discuss the practical application of the proposed methodology and conclude.

2. Methodology

Central to this study is the assumption that individual observed student outcomes, eg module grades, are not necessarily equal to ideal student outcomes, after controlling for individual student characteristics. The difference between ideal student outcomes and observed student outcomes is defined as the student attainment gap (u), and is a measure specific to each assessed student (i):

$$u_i = ideal\ outcomes_i - realised\ outcomes_i \qquad \text{(Eq. 1)}$$

This study assumes that good teaching practice should have as primary objective the minimisation of the student attainment gap; in other words to help students reach their maximum potential. This axiom should be relatively uncontroversial – various previous studies view the concept of teaching quality as a transformation process for the student that aims at enhancing their capabilities (Gibbs 2010; Harvey and Green 1993). In fact, the attainment gap is a measure similar to educational gain, which was previously proposed in the literature (Gibbs 2010). Educational gain is defined as the difference between student performance before and after the student’s learning experience; Gibbs suggests that this could be measured through a standardised assessment taken before the module takes place and after its completion. However, measuring educational gain at the module level in this manner is problematic, as it involves substantially increasing the assessment burden to the students.

The attainment gap measure proposed here is more realistic; a student’s performance after the learning takes place can be assessed through the normal summative assessment of the module itself. Assuming that the module is constructively aligned (Biggs 1996), students’ performance in the module’s assessment should provide a robust and parsimonious measure of student learning. The issue then becomes; how to arrive at a reasonable measure of ideal outcomes and how to control for student characteristics that might affect these benchmarks?

To address these issues, the study utilises the concept of the educational production function. Educational production functions have been widely used in the literature (for some earlier examples, see Levin 1974; Hanushek 1979); their aim is to provide quantitative evidence that links the performance of a student to educational inputs and student characteristics. Perelman and Santin (2011) provide a brief overview of the generalised theoretical model:

$$A_i = f(B_i, S_i, P_i, I_i) \qquad \text{(Model 1)}$$

, where, *A* is the learning outcome of student *i, B* is the student’s background, *S* are the educational (school) inputs, *P* is the peer group effect and lastly *I* is the student’s innate ability. These factors will thereafter be referred as ‘student characteristics’

As noted by Perelman and Santin (2011), in the majority of studies employing educational production functions, student performance is typically aggregated at the school/institution/regional level. Only a few studies use disaggregated student data directly in the analysis (see Johnes et al. 2017; De Witte and Lopez-Torres 2017 for a review) and none so far have examined student performance at the module level. However, the educational production function presented as Model 1 can easily be applied at the student/module level. In fact, the general nature of Model 1 is an advantage, since the model to be estimated should be specific to the module that is being assessed. For example, if all students in a given module receive the same educational inputs (eg same classrooms, laboratories or study rooms, library facilities, etc), there is no need to include variables specific to this category in the model.

Model 1 results in a function that quantifies the expected or 'average' individual student performance controlling for student characteristics. To assess the attainment gap however, the analysis requires a function that describes ideal performance, again on the basis of each individual student's characteristics. To achieve that, the analysis moves from the concept of an educational production function to an educational production frontier.

Educational production frontiers have been utilised numerous times in previous studies, although as noted earlier, the level of analysis is usually at the institution/district/regional level. They are based on concept of the 'traditional' production frontiers, which attempt to model a primal transformation function (ie a process that utilised certain inputs to derive certain outputs without recourse to input and output prices) while accounting for the fact that not all production is optimal, ie:

$$Y_i \leq f(X_i) \qquad \text{(Eq. 2)}$$

,where Y is a vector of outputs, X is a vector of inputs and i=1…n is the sample of producers that are being modelled. Eq. 2 can be converted to equality by introducing u_i, the element of technical inefficiency:

$$Y_i = f(X_i) - u_i \qquad \text{(Eq. 3)}$$

, where $u_i \geq 0$. Rearranging Eq.3, we have:

$$u_i = f(X_i) - Y_i \qquad \text{(Eq. 4)}$$

, which is very similar to Eq. 1, where Y_i represents realised outcomes and $f(X_i)$ represents ideal outcomes.

To derive $f(X_i)$, one can use any number of the various frontier estimation approaches suggested in the literature. In the field of education, the majority of research utilises non-parametric methods, such as Data Envelopment Analysis (DEA) and Free-Disposal Hulls (FDH); see table 8 in De Witte and Lopez-Torres (2017) for a breakdown of studies by estimation technique. This study instead utilises a parametric Stochastic Frontier Analysis (SFA) model in order to parameterise both $f(X_i)$ and derive the expectation of u_i. In general, each frontier estimation technique has its strengths and weaknesses. Non-parametric techniques generally require fewer assumptions (for the functional form of the transformation function and the distribution of the inefficiency term) relative to the parametric approaches and allow for both multiple inputs and outputs, while the parametric approaches traditionally allow for only a single output. The main strength of SFA, developed independently by Aigner, Lovell, and Schmidt (1977) and by Meeusen and van Den Broeck (1977), is that it explicitly allows for a stochastic element in educational production function model.

Under the ‘classical’ deterministic frontier approaches, such as the original models of DEA and FDH, any external event that has an impact on the transformation process, such as a student taken ill in the middle of the term and missing some lectures, but also any measurement error random error in the data used in the analysis (be it inputs, outputs or other contextual variables) would have a direct impact on the estimate of u_i. SFA attempts to account for those random effects in performance by directly including a stochastic element in the formulation of the (educational) production function.

This is important in the context of this analysis, since defining and estimating the process of student learning through an educational production function is a complex task (Worthington 2001; Johnes 2014). Regardless of the complexity of the models, there will always be factors that affect individual student performance that the model will not be able to account for, mainly due to lack of granular, student-specific data. For example, the mental state of the student throughout the duration of the module and especially during the assessment period can have a significant impact on the student's attainment but measuring and validating such a variable is very difficult to do in institutions with thousands of students and hundreds of modules. As such it is very beneficial to the analysis to incorporate a stochastic element to the model to acknowledge the limitations of the model's specification. By incorporating a stochastic element (ε_i), Eq. 3 evolves to:

$$Y_i = f(X_i) + \varepsilon_i - u_i \qquad \text{(Eq. 5)}$$

, where ε_i is the standard two-sided, normally distributed error term found in regression analysis, which is also assumed to be independently distributed of u_i.

Equation 5 cannot be estimated through Ordinary Least Squares regression (OLS); however, if u_i is independently distributed relative to the inputs, the OLS coefficients are statistically consistent except for the constant. Therefore, OLS can be used as a first step to estimate the slope parameters (coefficients of all the inputs) and then a second method can be used to estimate the constant and the two residual components, namely the stochastic element and technical inefficiency.

The SFA literature provides two main methods for the second step, Maximum Likelihood estimation (MLE) and the Method of Moments approach (MoM). For both approaches, an additional assumption about the distribution of the u_i term is required. The literature has proposed a number of possible distributional assumptions, but the most common distributions used in practical applications are the half-normal and the exponential[3].

It should be noted here that advances in non- and semi-parametric techniques resulted in formulations that can also account for a stochastic element in the modelling process, either indirectly (through bootstrapping DEA models (Simar and Wilson2000) or directly (Kuosmanen and Kortelainen 2011). However this comes at the cost of significantly increased complexity in the estimation and generally more stringent data requirements (ie large datasets). And although the ability of the SFA models to directly include the stochastic element in the estimation process is a major reason for their use in this application, there two other very important considerations that are relevant to this setting. The first is that the parametric approaches provide a clear description of the educational production function and allow for easy estimation of significance levels for the model coefficients; this can be very helpful in assessing the impact and significance of different student characteristics on attainment. The second consideration is more practical in nature; student characteristics often take the form of discrete variables (eg student gender, student funding status, etc) and non-parametric approaches do not allow for such variables to be included in a single model[4].

After selecting the distribution for u_i, MLE estimates the parameters of the production function in such a way so they provide the highest joint probability of observing the current sample, utilising the OLS parameters as a starting point. For a more in-depth description of the estimation process, see Kumbhakar and Lovell (2000). The end result of the process is an estimate of the expectation of u, E(u). It should be stressed here that the approach does not

[3] The prevalence of these distributions in practical analysis is mainly their ease of use; both are one-parameter distributions so if one moment of the distribution is known, all the other moments of interest can be derived analytically. For a comprehensive discussion on the possible distributions that can be used for the decomposition, see Greene (2008).

[4] Non-parametric approaches can accommodate discrete variables in the analysis by creating different model for each discrete group. However, modules/courses have usually relatively small sample sizes, which could result in some groups being too small. Additional complications arise when there are multiple discrete variables and the analysis wants to test how they interact.

provide direct estimates of the u_i term, since the term is unobserved. In 'traditional' SFA applications, E(u) is then used to generate an estimate of the conditional mean of inefficiency, E($u_i|\varepsilon_i$), which can then be used to generate estimates of technical inefficiency for all assessed units based on the distributional assumption for the term. However, one of the main criticisms of SFA is that these conditional estimates are statistically inconsistent in the cross-sectional setting[5] (Kumbhakar and Lovell 2000). A strength of the proposed approach here is to omit this last step and instead simply focus on the E(u) measure. As a reminder, although the unit of the assessment is the individual student, the analysis is not interested in individual estimates, but rather on assessing the overall attainment gap for the module. As such, only estimates of E(u) are needed, which SFA can consistently estimate, and the common issue in SFA regarding the statistical inconsistency of unit-specific estimates does not apply here.

The MoM approach also produces estimates of E(u) but using a different methodology. In simple terms, if u_i is present in the educational production function, the OLS regression should result in an error term (e_i) that is skewed to the left, since $e_i = \varepsilon_i - u_i$ and $\varepsilon_i \sim N(0, \sigma_v^2)$ and $u_i \geq 0$. The mean of the composed error term is by construction 0, but the information provided by its third moment (its skewness) together with the distributional assumption for *u* can be used to estimate the variance of the *u* term, σ_u^2. If it is assumed that $u_i \sim N^+(0, \sigma_u^2)$ (ie it is half-normally distributed), then:

$$\hat{\sigma}_u^2 = \left(\frac{m_3}{\sqrt{\frac{2}{\pi}}(1-\frac{4}{\pi})} \right)^{2/3} \qquad \text{(Eq. 6)}$$

, where m_3 is simply the third moment of the OLS composed error term, e_i. Since the half-normal distribution is a one-parameter distribution, $\hat{\sigma}_u^2$ can be used to derive E(u):

[5] Statistically consistent estimates of ui are possible in the panel setting, but only under the assumption that ui remains unchanged for the duration of the analysis, which is very restrictive in general and inappropriate in the setting of this study.

$$E(u) = \left(\sqrt{\frac{2}{\pi}}\right) \hat{\sigma}_u \qquad \text{(Eq. 7)}$$

If it is assumed that $u_i \sim Exp(u)$ (ie it is exponentially distributed), then:

$$\hat{\sigma}_u^2 = \left(-\frac{m_3}{2}\right)^{2/3} \qquad \text{(Eq. 8)}$$

, and once again $\hat{\sigma}_u^2$ can be used to derive E(u):

$$E(u) = \ \sigma_u \qquad \text{(Eq. 9)}$$

Regardless of the estimation methodology, the end result is an estimate of E(u), ie the expected (average) value of the attainment gap for the whole student cohort. The functional form used for the educational production function will determine the terms that E(u) is expressed. Since Eq.1 is a linear relationship, a linear functional form for the education production frontier is appropriate in this application. In this case, E(u) will have the same scale as the dependent variable.[6]

An estimate of E(u) for a single academic year and a single module is of limited value on its own, since there is nothing to compare it with. Instead, the estimate of interest here would be the rate of change of E(u), from one year to the next for the same module, ie:

$$\mathit{\Delta} E(u) = \frac{E(u)_t}{E(u)_{t-1}} - 1 \qquad \text{(Eq. 10)}$$

Positive values of $\mathit{\Delta} E(u)$ show a decrease in the average attainment gap, while negative values show an increase. As such the analysis should utilise a series of models, one for each student cohort that sits the specific module, and track how the attainment gap changes over time. Alternatively, the proposed methodology can be used to evaluate changes in the

[6] Note that Eq.1 can be formulated so that: $u_i = \frac{realised\ outcomes_i}{ideal\ outcomes_i}$. In this case, the educational production function should be expressed using a logarithmic functional form, such as the Cobb-Douglas or the translog and u_i is expressed as a multiplier that converts realised outcomes to ideal outcomes of vice versa.

structure and/or delivery of a module, by comparing the estimated average attainment gap before and after such changes. The application below demonstrates such a case.

3. Case study

This study was initially motivated by my experience with teaching basic Mathematics and Statistics to first year undergraduate Business School students at Aston University, UK. It is well documented that teaching mathematics and/or statistics to students that are not studying for a science degree or engineering is a challenging task, possibly due to a general fear of the quantitative nature of the material to students that study for a non-quantitative or mixed degree (Cybinski and Selvanathan 2005; Onwuegbuzie and Wilson 2003). The current iteration of the module is called Foundations in Business Analytics (FBA) and has recently undergone substantial changes in terms of its delivery; the teaching team was reduced to just three experienced academics with no teaching assistants, the teaching groups for the practical sessions grew from approximately 20 to 100 students per session, weekly live, on-line revision seminars (webinars) were introduced and the user interface of the on-line teaching material available to the students was modernised. These substantial changes in the method of delivery make this a very suitable case study for the proposed application.

The first step of the analysis is to parametrise the educational production function. As a reminder, the general form of the educational production function takes into account four general categories of educational inputs: the student’s background, the school-related inputs, the peer group effect and lastly the student’s innate ability. In this application, all students in each cohort benefited from the same school-related inputs; there were no specific groups of students that had access to more school-related resources relative to others and as such this general category is not applicable here. Peer group effects are similarly not included, since the module does not have a formal peer-group allocation system for studying and has no group-related assessments. It is very likely that some students will form study groups with their peers but data on such practices are not available. This results in the following general model:

$$A_i = f(P_i, Sp_i, I_i, D_i, Rs_i) \qquad \text{Model 2}$$

, where P is the student's prior attainment, Sp is whether the student has declared any special learning needs or any learning disability, D is the demographic characteristics of the student and Rs is whether the student is repeating the module.

In more detail:

Learning outcomes (A): FBA is designed such that the learning outcomes are very closely aligned with the module's assessment practices. As such, the most appropriate way to quantify by how much students have achieved the stated learning outcomes is through their final module mark. Therefore, this study uses the final module mark (*FBAmark*) to approximate student achievement.

Prior attainment (P): The nature of the subject matter in FBA is 'cumulative' and as such, the prior level of attainment in Mathematics and Statistics subjects is likely to be a very important factor in explaining student performance. The hypothesis is that students that have a solid foundation in Mathematics will find it easier to deal with the concepts introduced in this module, due to prior experience with the subject and reduced subject-related learning anxiety (Onwuegbuzie and Wilson 2003). At the same time, students that had no further engagement in any quantitative subjects after completing their secondary education (ie after their GCSEs), are likely to find it more difficult to adjust to the demands of a heavily quantitative module. This study uses a number of indicators for prior attainment, all included as separate indicator variables:

1. A-level qualification[7] in Mathematics and/or related subjects, awarded with C or above. (*AL_Maths_qual*)
2. A-level qualification in Statistics and/or related subjects, awarded with C or above. (*AL_Stats_qual*)
3. Other Mathematics or Statistics subject qualification from a recognised body (eg Functional Skills qualification, International Baccalaureate) or an AS-level qualification[8] in Mathematics and/or related subjects (while not having an A-level in the same subject) or an A-level qualification in Mathematics and/or related subjects, awarded with D. (*Other_qual*)
4. BTEC (Business & Technology Education Council, Edexcel) qualification, without any A-level qualifications. From past experience teaching in the subject area, some BTEC students demonstrate low levels of confidence in their mathematical ability and sometimes there are significant gaps in their understanding of the basics of the subject matter. (*BTEC*)
5. ‘Access to HE Diploma’ (AHE) qualification, without any A-level qualifications. Students with such a qualification are usually mature students who have been out of education for some time. (*Access*)

Special learning needs and/or learning disabilities (Sp): The University offers specialised support for students with learning disabilities or special learning needs (such students use the *DANU* identifier) and often makes separate, individual examination arrangements. It would be interesting to examine whether declared learning disabilities/special learning needs have an effect on student performance, after controlling for other individual student characteristics.

[7] In the UK, students typically study for A-levels, a subject-based qualification, after they completed their secondary education and before applying for a university place. All universities require certain A-levels achieved or equivalent qualifications.

[8] A-levels are typical split into two parts, each assessed separately. Students that qualify for the first part only are awarded an AS-level qualification.

Demographic characteristics (D): Two elements are included here, gender and student origin. Gender is included because there is evidence that suggests male students tend to perform better in mathematic-related subjects at the end of their secondary education (from the PISA 2012 study, OECD (2012)). Student origin was included to test whether there are any significant differences in performance across three broad student groups, those that that have the UK as their home origin, students that come from other EU countries and non-EU overseas students. The model uses the UK-origin group as the base.

Innate academic ability (I): Although this is probably the most important variable in explaining student performance, it is also one of the most difficult to measure. A good proxy in this case would be student performance in a standardised entry assessment, similar to a SAT score. However, there is no similar widespread standardised assessment framework in the UK for students that finish secondary education. As such, this study uses the average grade that students have achieved in all first year undergraduate modules of their course, except FBA, (*AvMark*) as a proxy[9].

Resit Exam (Rs): This is an indicator of whether the student has repeated the module or has passed the module in the resit assessment.

The data used in this study were sourced directly from the University registry system. After the data was cleared of anomalies (students with missing values in some characteristics and instances where specific students could not be matched to available grades), the sample size for the first (2013/14) cohort, before the changes to the module were implemented, was 663 and 658 students for the second (2014/15) cohort respectively.

[9] Other indicators of academic performance could also be appropriate here, such as graduation mark. This study adopts the AvMark indicator to preserve the immediacy of the analysis (ie the analysis can be undertaken immediately after the module has concluded and assignments are marked). It is also arguable that a single-year average is more relevant in this case, as it more starkly captures transient effects on student performance specific to the year in question that the model does not directly measure (due to lack of data on individual student circumstances).

An overview of the data for the two years available for this study is presented in Table 1.

Table 1: Descriptive statistics for the inputs and outputs

	2013/14 (663 students)			2014/15 (658 students)		
	Yes	No	Frequency of Yes	Yes	No	Frequency of Yes
AL_Maths_qual	66	597	9.95%	59	599	8.97%
Other_qual	36	627	5.43%	42	616	6.38%
AL_Stats_qual	3	660	0.45%	2	656	0.30%
Gender (Male)	380	283	57.32%	386	272	58.66%
Resit	31	632	4.68%	18	640	2.74%
Home	516	147	77.83%	523	135	79.48%
EU	47	616	7.09%	34	624	5.17%
Overs	100	563	15.08%	101	557	15.35%
DANU	23	640	3.47%	22	636	3.34%
BTEC	178	485	26.85%	199	459	30.24%
Access	72	591	10.86%	75	583	11.40%
	Mean	Standard Deviation	Maximum	Mean	Standard Deviation	Maximum
FBAmark	44.4	21.1	93.1	44.6	19.4	93.6
AvMark	51.9	15.4	80.8	51.5	14.5	82.4

The first model of this analysis is a linear multiple OLS regression to examine the impact of the various student characteristics on student performance:

$$A_i = \text{AvMark} + \text{AL\_Maths\_qual}_i + \text{Other\_qual}_i + \text{AL\_Stats\_qual}_i + \text{BTEC}_i + \text{Access}_i + \text{Gender}_i + \text{EU}_i + \text{Overs}_i + \text{DANU}_i + \text{Resit}_i \qquad \text{Model 3}$$

Model 3 was estimated for both 2013/14 and 2014/15 separately using general-to-specific[10]. All variables are indicator (dummy) variables, with the exception of the dependent variable (student attainment, given by the mark achieved by each individual student for the module)

[10] General-to-specific iterates on the general model by removing a single insignificant variable at each stage, re-evaluating the model and repeating the variable removal step until all remaining variables are statistically significant (moving from a general model specification to a specific one that includes only statistically significant variables). If there are multiple insignificant variables in a given step, the process selects the variable with the highest p-value for removal.

and the AvMark variable, which is a proxy for innate academic ability. The results of the analysis are presented in table 2:

Table 2: Model 3 results

	2013/14			2014/15	
	Coeff.	Std. Err.		Coeff.	Std. Err.
AvMark	0.89 **	0.03	**AvMark**	0.92**	0.03
AL_Maths_qual	16.43 **	1.73	**AL_Maths_qual**	13.48**	1.62
Other_qual	6.09 **	2.21	**Other_qual**	4.32*	1.88
AL_Stats_qual	16.77 *	7.44			
Gender	4.98 **	1.01	**Gender**	2.68**	0.92
Resit	-15.99 **	2.37	**Resit**	-17.10**	2.77
EU	10.07 **	2.01	**EU**	8.63**	2.07
Overs	5.03 **	1.47	**Overs**	3.43**	1.32
BTEC	-3.69 **	1.27	**BTEC**	-3.88**	1.12
Constant	-6.45 **	2.18	**Constant**	-5.16*	2.03
Model Statistics			**Model Statistics**		
Adj. R^2	0.635		Adj. R^2	0.650	
RMSE	12.773		RMSE	11.486	

Note: * denotes significance at the 5% confidence level, **denotes significance at the 1% confidence level. DANU and Access variables were statistically insignificant and were dropped from the analysis.

As expected, students with A-levels in Mathematics or Statistics tend to do significantly better than the average, controlling for all other significant factors. Students with other Maths related qualifications do better on average, although not as well as students with A-levels. This is strong evidence to support the view of the 'cumulative' nature of the subject matter.

Male students perform better relative to female students, although this effect is quite small and decreasing over time. More important is the clear distinction between students that come from different educational systems. EU-origin students are clearly doing better than Home-origin students, with a difference of approximately 9 to 10 percentage points. This might be due to the fact that prior mathematics-related qualifications for EU students are not captured sufficiently well by the model; another possible explanation is that the majority of EU countries place higher emphasis in mathematics-related subjects in their secondary education curricula relative to the UK[11]. Overseas students are also doing significantly better than Home students, though the difference in average performance is smaller (approximately 3 to 5 percentage points).

BTEC students perform worse on average, even though the model accounts for the fact that BTEC students tend to perform poorly in all 1st year modules (the difference in average scores of BTEC to non-BTEC students in all 1st year modules is approximately 11 percentage points for both years of the analysis).

With regards to the statistical properties of Model 3, the 2013/14 specification has a slightly lower adj. R^2 than the 2014/15 specification but the difference is marginal. The 2013/14 passes the Breusch-Pagan / Cook-Weisberg test for heteroskedasticity, but fails both the

[11] Most EU students are taught Mathematics to up their last year of secondary education, while UK students that don't study a quantitative subject for their A-levels end their engagement with the subject when they finish their GCSE's, which is usually at 16 years of age. The hypothesis that the quality of Mathematics education in UK at secondary level is somewhat lagging behind its EU counterparts is unlikely to be valid, as the PISA results for 15 year olds finds that UK student performance in Mathematics is very close to the average of the OECD countries (OECD 2012).

White's test and the RESET test for misspecification. On the other hand, the 2014/15 model passes all three diagnostic tests successfully, in all their permutations. Despite these apparent differences, both specifications are very similar when it comes to the significance of their coefficients and the magnitude of their effects. Both specifications find the same variables to be statistically significant, with AL_Stats_qual being the sole exception. It should be noted here however that only 3 students in 2013/14 and only 2 students in 2014/15 had an A-level in Statistics, so the general instability with regards to that variable is not unexpected.

The remaining models for this analysis are based on Model 3 with the student attainment gap (u_i) included in the model specification; they are estimated using both MLE and MoM approaches[12]:

$$A_i = \text{AvMark} + \text{AL\_Maths\_qual}_i + \text{Other\_qual}_i + \text{AL\_Stats\_qual}_i + \text{BTEC}_i + \text{Gender}_i + \text{EU}_i + \text{Overs}_i + \text{Resit}_i - u_i$$

Model 4-i (2013/14 specification)

$$A_i = \text{AvMark} + \text{AL\_Maths\_qual}_i + \text{Other\_qual}_i + \text{BTEC}_i + \text{Gender}_i + \text{EU}_i + \text{Overs}_i + \text{Resit}_i - u_i$$

Model 4-i (2014/15 specification)

In both specifications of model 4 the composed error term had a negative skew, which is a strong indication that there is a gap between optimal and observed performance. For the MLE decomposition, chi-squared tests for the significance of σ_u rejects the null hypothesis that σ_u is equal to zero at less than a 1% significance level. The estimates of the expected

[12] Models 4-I and 4-ii include the exact same variables as those used the more general model 3 and adopt the same functional form specification (linear). In the case of MoM, the model coefficients are also identical by construction. For MLE, there are some very small differences in the coefficients between Model 3 and Model 4, due to the change from OLS estimation to Maximum Likelihood estimation. However, the same set of variables that were statistically significant in under Model 3 are also statistically significant under Model 4

performance gap (E(u)) under both exponential and half-normal assumptions for u_i are given in tables 3; the half-normal assumption resulted in an infeasible solution for the MLE specification in 2013/14.

Table 3a: Estimates of the performance gap derived from Model 4, MLE specification

	2013/14		**2014/15**		$\boldsymbol{\Delta E(u)}$
	E(u)	Significance	E(u)	Significance	
Half-normal	Not feasible		10.00	0.000	N/A
Exponential	16.12	0.000	6.65	0.000	**58.7%**

Note: Significance relates to the p-value of the chi-squared test for σ_u =0.

Table 3b: Estimates of the performance gap derived from Model 4, MoM specification

	2013/14		**2014/15**		$\boldsymbol{\Delta E(u)}$
	E(u)	Significance	E(u)	Significance	
Half-normal	31.85	0.000	10.49	0.000	**67.1%**
Exponential	16.97	0.000	6.28	0.000	**63%**

Note: Significance relates to the p-value of the Jarque-Bera, sample size-adjusted test for zero skewness of residuals.

Noteworthy in the above results is that both specifications produce very similar estimates of the attainment gap. Also of note is that although the two distributional assumptions produce quite different estimates of the average attainment gap, the estimate of the change in attainment gap over the two cohorts is very similar under the exponential assumption, ie the assumption that produces a full set of results.

Overall, the average performance gap has decreased by approximately 59% to 67%, which is a very significant improvement in student performance. It is interesting to compare this result with more commonly employed measures of teaching quality, namely student feedback scores and unadjusted student marks. As table 1 demonstrates, average student marks improved in 2014/15 relative to the previous year, but only slightly so (0.2 percentage points). Student feedback scores were also very similar between the cohorts; in fact, the feedback scores in the majority of the survey questions were lower than the previous year and as a result, the overall module score fell by approximately 3%. On the other hand, the teaching team was of the view that the structure, organisation and delivery of the module in the 2014/15 cohort had significantly improved.

It is worth trying to understand how the approach produced such a large estimate of the change in student attainment gap between the cohorts and provide possible links to the actual changes implemented in practice. From a technical perspective, the cohort-specific estimates are a function of model fit, ie how well the model describes the data, and the similarity of the model's residuals to a variable with a combined normal/half-normal (or normal/exponential) distribution. For the first factor, the higher the fit, the smaller the attainment gap, simply because there is less unexplained variation in the model to be assigned to the attainment gap; the intuition behind this was discussed in section 2. In this case study, model fit has indeed improved in the second cohort, as demonstrated by the RMSE values in Table [2]. It should be noted that the overall variation in the dependent variable, FBA mark, is only very slightly smaller in the second cohort, and as such the reduced RMSE values are indeed due to the improvement in overall model fit. In the second cohort, there are also fewer students with large residuals, which allows this cohort's residual distribution to fit more easily to the normal/exponential distribution of the theoretical model. These two factors combined are likely the main drivers behind the reported results.

From a pedagogical perspective, the changes implemented to the module's delivery aimed primarily at reducing the variance of the student experience with the module. The reduction of the teaching team to just three experienced teachers meant that all students would receive similar quality of instruction, which is especially relevant to the practical sessions that are critical for quantitative modules. The modernisation of the delivery of the on-line material lowered the barrier of accessing said material. The introduction of webinars and extra sessions for students that needed more help provided more opportunities to students to engage with the module. It is quite plausible that these changes taken together should reduce the unexplained variability in student attainment. As example, let's assume that we have two students with the same characteristics; one is being taught by an inexperienced teaching assistant and another by an experienced lecturer. It is quite likely that the second student will achieve a higher mark, but since the model does not distinguish whether a student is taught by experienced or inexperienced teachers, it will assign this difference in performance to unexplained variation and thus the model fit will suffer. When all students are taught be experienced teachers, the source of this particular variation disappears and thus model fit improves, relative to the previous state. It should be noted here that there might be other hidden factors that might have affected model fit in this particular case study; omitted variable biases are not uncommon in educational production functions. However, the approach adopted here made every effort to mitigate such potential biases by relying on an intuitive theoretical model, utilising all available data and generating separate models for the two cohorts (thus avoiding imposing the assumption of coefficient stability between the cohorts).

Although anecdotal, the above suggests that the proposed approach provides insights on the issue of teaching quality not captured in more ‘traditional’ methods.

4. Implementation issues and conclusions

The main contribution of this study is a new model for quantitatively assessing the student attainment gap and how this can be used to derive an estimate in the change of teaching quality, utilising some basic principles from Stochastic Frontier analysis models. It assumes that the main goal of teaching is to help students realise their full potential and that a student’s grade is a good indicator of learning outcomes achieved (ie the learning outcomes and the assessment are constructively aligned). If the above holds, a module -specific educational production frontier can be constructed based on the data of individual student characteristics in a given student cohort.

The resulting analysis has a number of strengths: it is outcomes-based, avoiding the issues and controversy surrounding student survey-based measures. It is also assessing outcomes controlling for student characteristics that may affect such outcomes; as such it is much more robust than examining simple average cohort grades or pass rates, which is especially relevant when student entry requirements are fluid from one cohort to the next. Additionally, because the attainment gap measure is an aggregate measure, there are no issues with the inherent statistically inconsistency of individual SFA-derived performance scores. Lastly, the proposed model is relatively easy to implement, given a robust educational production function; MLE is available in the majority of statistical software packages and the MoM approach only requires software that can estimate simple OLS models.

For researchers interested in implementing the proposed model, some issues to bear in mind:

Sample size: The decomposition of the composed error term of the original model benefits greatly from large sample sizes (Kumbakhar, Wang and Horncastle 2015; Behr and Tente 2008). Although the decomposition can be attempted when sample sizes are small, the accuracy of the resulting E(u) is not guaranteed. Nevertheless, for situations where the stochastic component of the composed error term is likely to be important, as is the case in educational production functions, simulation evidence presented in Behr and Tente (2008) suggest that model accuracy improves by approximately 25% to 60%, depending on the prevalence of the stochastic element relative to u_i, when moving from 25 observations to 1000 observations and using the MoM specification. In practical terms, the proposed methodology is not suitable for modules with small student numbers; as a rule of thumb, it is suggested that the model has at least 30 degrees of freedom available for the decomposition (ie student numbers minus the number of student characteristics included in the educational production function should exceed 30).

MLE or MoM: Related to the above, both estimation approaches provide statistically consistent estimates of the average attainment gap. Simulation studies (Coelli 1995; Behr and Tente 2008) have shown that MoM provides greater accuracy in small and medium sized samples when the attainment gap is not strongly dominating the stochastic element. In practical terms, the educational production function is likely to be ‘noisy’, ie display a relatively high root mean square error and have a modest fit (as an example, the adjusted R^2 in the models of this case study ranges from between 60 to 65%); as such, it is likely that the variance of stochastic element should be comparable to the attainment gap and not orders of magnitude smaller. Module class sizes vary, but usually student cohorts are expected to be about 50 to 100 students. Given the above, MoM is likely to be more suitable in most applications of the approach.

Reliance on the educational production function: The starting point of the analysis is the educational production function and how this function is parameterised can have a large impact on the attainment gap estimates. As is the case with traditional production frontiers, the overall fit of the model will be correlated with the level of the estimated attainment gap; as such, higher model root mean squared errors will result in higher attainment gap estimates. As previously discussed, the educational production function is trying to model a very complex process (student learning) and it is very likely that not all factors that affect this process will be observable to the researchers (for example, student innate academic ability is very difficult to measure). Therefore, the resulting models are expected to display relatively modest levels of fit, as demonstrated in the case study. It should be noted however that this is not a severe detriment to the analysis, given that the measure of interest is changes in attainment gap; since the cohort-specific estimates come from two models with likely similar statistical characteristics (as demonstrated in the case study), the baseline attainment gap estimates from each individual cohort are more likely to be broadly comparable. Nevertheless, researchers should strive to achieve the best fit possible, allowing for the data available to the analysis.

The module-specific nature of the analysis: The proposed methodology is not appropriate to benchmark teaching quality across different modules. As mentioned above, the estimated attainment gap is highly correlated with model fit and some modules will display educational production functions with higher fit, simply because there is lower variation in student outcomes (for example, modules with simple learning objectives where students have already been exposed to the learning material at a prior stage). For these modules, the estimated attainment gap will be modest, relative to a more challenging module. The difference in the student attainment gap between such modules is more a function of the module learning material and objectives and as such it cannot be assigned to differences in teaching quality.

Changes in module assessments between reviews: Related to the point above, changes in the method of assessment/difficulty of assessment between cohorts are not likely to significantly influence the accuracy of the attainment gap estimates. The approach proposed here is robust to assessment changes when those changes make the assessment more or less challenging equally across all students (the effect will be captured by the constant in the regression). The approach is also robust to assessment changes that can be captured by student characteristics (eg by providing more time to complete exams for students with special learning needs, providing additional support to students with less prior exposure to the subject). Changes that will adversely affect the estimates are those that reduce the variability of student outcomes after controlling for student characteristics. For example, artificially increasing the marks of the lowest scoring students, while keeping other marks as is.

To conclude, this study hopefully demonstrates some of the strengths quantitative analysis can bring to research in education and provides an additional tool in outcome-based assessment of module/teaching quality.

References

Aigner, D., Lovell, C. A. L., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. Journal of Econometrics, 6(1), 21-37.

Barra, C. & Zotti, R. (2016). Managerial Efficiency in Higher Education Using Individual Versus Aggregate Level Data. Does the choice of Decision Making Units Count?, Managerial and Decision Economics, John Wiley & Sons, Ltd., 37(2), 106-126.

Behr, A. & Tente, S. (2008). Stochastic Frontier Analysis by Means of Maximum Likelihood and the Method of Moments. Bundesbank Series 2 Discussion Paper No. 2008,19. Available at SSRN: https://ssrn.com/abstract=2794022

Biggs, J.B. (1996). Enhancing teaching through constructive alignment. Higher Education, 32(3), 347-364

Brusoni, M. (et al.) (eds.) (2014). The Concept of Excellence in Higher Education. Brussels: European Association for Quality Assurance in Higher Education

Coelli, T. J. (1995). Estimators and hypothesis tests for a stochastic frontier function: A Monte Carlo analysis, The Journal of Productivity Analysis 6, 247–268.

Cybinski, P., & Selvanathan, S. (2005). Learning Experience and Learning Effectiveness in Undergraduate Statistics: Modeling Performance in Traditional and Flexible Learning Environments. Decision Sciences Journal of Innovative Education, 3(2), 251-271.

Bernstein, J. B. (2008). Peer Review and Evaluation of the Intellectual Work of Teaching, Change: The Magazine of Higher Learning, 40:2, 48-51, DOI: 10.3200/CHNG.40.2.48-51

De Witte, K & López-Torres, L. (2017). Efficiency in Education: A Review of Literature and a Way Forward (March 2017). Journal of the Operational Research Society, 68 (4)

Gibbs, G. (2010). Dimensions of quality. York, UK: The Higher Education Academy.

Greene, W. (2008). “The econometric approach to efficiency analysis” in Fried, H. O., Lovell, C. K., & Schmidt, S. S. (eds) “The measurement of productive efficiency and productivity growth”. Oxford University Press, USA.

Gunn, V., & Fisk, A. (2013). Considering Teaching Excellence in Higher Education: 2007–2013: A Literature Review Since the CHERI Report 2007. York: Higher Education Academy.

Hammersley-Fletcher, L. & Orsmond, P. (2004). Evaluating our peers: is peer observation a meaningful process?, Studies in Higher Education, 29:4, 489-503, DOI: 10.1080/0307507042000236380

Hanushek, E., A. & Wößmann, L. (2010). “Education and Economic Growth”. In: Penelope Peterson, Eva Baker, Barry McGaw, (Editors) “International Encyclopedia of Education” Volume 2, 245-252. Oxford: Elsevier.

Hanushek, E.A. (1979). ‘Conceptual and empirical issues in the estimation of educational production functions’, Journal of Human Resources, 14 (3), 351–88.

Harvey, L. & Green, D. (1993). Defining Quality. Assessment & Evaluation in Higher Education,18(1), 9 –34

Hinchey, P. H. (2010). Getting Teacher Assessment Right: What Policymakers Can Learn from Research. National Education Policy Center. http://nepc.colorado.edu/publication/getting-teacher-assessment-right.pdf

Johnes, J. (2014). Operational Research in education. European Journal of Operational Research. 243(3), 683-696.

Johnes, J., Portela M. & Thanassoulis, E. (2017). Efficiency in education, Journal of the Operational Research Society, 68(4), 331-338, DOI: 10.1057/s41274-016-0109-z

Kumbhakar, S. C., & Lovell, C. K. (2000). Stochastic frontier analysis. Cambridge University Press.

Kumbhakar, S., Wang, H., & Horncastle, A. (2015). A Practitioner's Guide to Stochastic Frontier Analysis Using Stata. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139342070

Kuosmanen, T., & Kortelainen, M. (2011). Stochastic Non-Smooth Envelopment of Data: Semi-Parametric Frontier Estimation Subject to Shape Constraints. Journal of Productivity Analysis, 35(2).

Land, R & Gordon, G (2015). Teaching Excellence Initiatives: Modalities and Operational Factors. York: HEA.https://www.heacademy.ac.uk/resource/teaching-excellence- initiatives-modalities-and-operational-factors

Levin, H. M. (1974). Measuring efficiency in educational production. Public Finance Quarterly, 2, 3-24

Meeusen, W., & van Den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. International Economic Review, 18(2), 435-444.

OECD (2013), PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy, OECD Publishing.

Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics Anxiety: Nature, aetiology, antecedents, effects, and treatments--a comprehensive review of the literature. Teaching in Higher Education, 8(2), 195-209.

Perelman, S., & Santin, D. (2011). Measuring educational efficiency at student level with parametric stochastic distance functions: an application to Spanish PISA results. Education Economics, 19(1), 29-49.

Simar, L. & Wilson, P. W. (2000) A general methodology for bootstrapping in non-parametric frontier models, Journal of Applied Statistics, 27(6), 779-802, DOI: 10.1080/02664760050081951

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching the state of the art. Review of Educational Research,83(4), 598-642.

Worthington, A.C. (2001). An empirical survey of frontier efficiency measurement techniques in education, Education Economics, 9(3), 245-268.