The quality and efficiency of public service delivery in the UK and China

Regional Studies

ISSN: 0034-3404 (Print) 1360-0591 (Online) Journal homepage: http://www.tandfonline.com/loi/cres20

The quality and efficiency of public service delivery in the UK and China

Minyan Zhu & Antonio Peyrache

To cite this article: Minyan Zhu & Antonio Peyrache (2017) The quality and efficiency of public service delivery in the UK and China, Regional Studies, 51:2, 285-296, DOI: 10.1080/00343404.2015.1080992

To link to this article: http://dx.doi.org/10.1080/00343404.2015.1080992

REGIONAL STUDIES, 2017
VOL. 51, NO. 2, 285–296
http://dx.doi.org/10.1080/00343404.2015.1080992

Routledge
Taylor & Francis Group

Regional Studies Association

The quality and efficiency of public service delivery in the UK and China

Minyan Zhu[a] and Antonio Peyrache[b]

ABSTRACT

The quality and efficiency of public service delivery in the UK and China. *Regional Studies*. This paper examines the efficiency of public service delivery at a regional level in both the UK and China using a method based on data envelopment analysis (DEA) that measures aggregate country-level inefficiency. This country-level inefficiency is then decomposed into three components: (1) lack of best practices at a regional level; (2) quality of the public service delivery; and (3) potential efficiency gains realizable via reallocation of expenditure across regions. The empirical results indicate that most UK inefficiency comes from the reallocation effect, while most Chinese inefficiency is attributable to lack of best practices; quality explains more of the expenditure variations in the UK relative to China. The paper speculates about fiscal (de)centralization as a possible explanation for such differences.

KEYWORDS

data envelopment analysis (DEA); directional distance function (DDF); fiscal decentralization; UK; China; public service quality

摘要

英国及中国的公共服务供给素质及效率. *区域研究*。本文运用根据测量国家层级总体非效率的数据包络分析（DEA）研究方法，检视英国和中国在区域层级的公共服务供给效率。此般国家层级非效率，接着被分解为三种构成要素：（1）缺乏区域层级的最佳实践；（2）公共服务供给的素质；以及（3）透过重分配各区域的支出可实现的潜在效率增加。经验结果指出，英国的非效率，多半来自于重分配的效果，而中国的非效率，则可归因于缺乏最佳的实践；与中国相较之下，素质更能解释英国的支出变异。本文推断财政（去）中心化作为此般差异的可能解释。

关键词

数据包络分析(DEA); 方向距离函数; 财政地方分权; 英国; 中国; 公共服务素质

RÉSUMÉ

La qualité et l'efficacité des prestations de services publics au R-U et en Chine. *Regional Studies*. À partir d'une méthode fondée sur l'analyse par enveloppement de données (DEA) qui mesure l'inefficacité globale à l'échelle nationale, ce présent article cherche à examiner l'efficacité des prestations de services publics au niveau régional à la fois au R-U et en Chine. Il s'ensuit une décomposition de cette inefficacité globale à l'échelle nationale en trois composantes: (1) le manque de pratiques optimales au niveau régional; (2) la qualité des prestations de services publics; et (3) les gains d'efficacité éventuels qui sont à réaliser par moyen d'une redistribution des dépenses à travers les régions. Les résultats empiriques indiquent que la plupart de l'inefficacité au R-U provient de l'effet de redistribution, tandis que la plupart de l'inefficacité en Chine s'explique plutôt par le manque de pratiques optimales; la qualité explique plus de la variation des dépenses au R-U par rapport à la Chine. L'article s'interroge sur la (dé)centralisation fiscale comme facteur explicatif éventuel de telles différences.

MOTS-CLÉS

analyse par enveloppement de données (DEA); fonction de distance directionnelle; décentralisation fiscale; R-U; Chine; qualité des services publics

CONTACT

[a] minyan.zhu@nottingham.ac.uk
Business School, University of Nottingham, Nottingham, UK.
[b] **(Corresponding author)** a.peyrache@uq.edu.au
Centre for Efficiency and Productivity Analysis (CEPA), School of Economics, The University of Queensland, Brisbane, QLD, Australia.

ZUSAMMENFASSUNG

Qualität und Effizienz bei der Umsetzung öffentlicher Dienstleistungen in Großbritannien und China. *Regional Studies*. In diesem Beitrag untersuchen wir die Effizienz bei der Umsetzung von öffentlichen Dienstleistungen auf kommunaler Ebene in Großbritannien und China unter Verwendung einer auf der Dateneinhüllanalyse (DEA) basierenden Methode zur Messung der gesamten Ineffizienz auf Landesebene. Anschließend wird diese Ineffizienz auf Landesebene auf drei Komponenten aufgegliedert: (1) Mangel an besten Praktiken auf regionaler Ebene, (2) Qualität der Umsetzung öffentlicher Dienstleistungen und (3) potenzielle Effizienzsteigerungen durch eine Neuaufteilung der Ausgaben zwischen den Regionen. Aus den empirischen Ergebnissen geht hervor, dass die meiste Ineffizienz in Großbritannien durch den Neuaufteilungseffekt verursacht wird, während die meiste Ineffizienz in China auf den Mangel an besten Praktiken zurückzuführen ist; die Qualität ist in Großbritannien stärker für Ausgabenschwankungen verantwortlich als in China. Wir spekulieren über eine fiskale (De)zentralisierung als mögliche Erklärung dieser Unterschiede.

SCHLÜSSELWÖRTER

Dateneinhüllanalyse (DEA); Direktionale Distanzfunktion; Fiskale Dezentralisierung; Großbritannien; China; Qualität öffentlicher Dienstleistungen

RESUMEN

Calidad y eficiencia de la prestación de servicios públicos en el Reino Unido y China. *Regional Studies*. En este artículo analizamos la eficiencia de la prestación de servicios públicos en un ámbito regional tanto en el Reino Unido como en China utilizando un método basado en el análisis envolvente de datos (DEA) que mide la ineficiencia global por país. Luego desglosamos esta ineficiencia por país en tres componentes: (1) falta de mejores prácticas a nivel regional; (2) calidad de la prestación de servicios públicos; y (3) posible aumento de la eficiencia a través de la redistribución de los gastos en las regiones. Los resultados empíricos indican que la principal ineficiencia británica procede del efecto de redistribución, mientras que la ineficiencia china se atribuye a la falta de mejores prácticas; las variaciones de los gastos en el Reino Unido se deben más a menudo a cuestiones de calidad que en China. En este artículo especulamos que la (des)centralización fiscal puede ser el motivo de estas diferencias.

PALABRAS CLAVES

análisis envolvente de datos (AED); función de la distancia direccional; descentralización fiscal; Reino Unido; China; calidad de los servicios públicos

JEL C44, R10, R50

HISTORY Received February 2014; in revised form July 2015

INTRODUCTION

In *The Wealth of Nations* (1776), Adam Smith stated one of the duties of the sovereign or commonwealth as

> that of erecting and maintaining those public institutions and those public works, which are useful but not capable of bringing in a profit to individuals. These are chiefly institutions for facilitating commerce and promoting instruction of the people.
>
> (Smith/Cannen, 1776/1904, bk V, ch. I, p. 244)

This paper examines the efficiency of public service delivery in two different countries, the UK and China, with two main innovations, one methodological and one empirical.

First, the quality of public service delivery is explicitly modelled. The quality of public service is becoming increasingly important for the general public and consequently for policy-makers. This is in line with the expectation that with rising living standards, demand will tend to shift towards higher quality services rather than larger quantities of low quality services (e.g., parents expect to receive a better education for their children, rather than expecting to 'park' them in school for longer hours). By taking into account the quality of public service delivery, an overall indicator (at both regional and national levels) which completely quantifies the observed trade-off between quality attributes and quantity outputs of public service delivery is defined. The most direct way of interpreting the indicator proposed here is to think of it as the overall loss in output quantity expansion (given input) that the government has to incur in order to attain the observed level of quality. The benchmark for such a comparison would be a baseline (zero) quality outcome which returns the highest possible output quantity expansion (the benchmark used here would be, for example, 'parking' children in school for long hours without providing any educational outcome). In other words, the indicator can measure the quality of public service in terms of the output quantity expansion that has to be traded off in order to attain the given level of quality (given input).

Once the quality/quantity trade-off has been accounted for, an aggregate measure of public service delivery inefficiency at the country level is defined as opposed to the

same measure computed at regional level. To achieve this a recent method (Peyrache, 2013) is adopted and modified in order to account for quality attributes. The measure of inefficiency at the country level is taken and shown that it is possible to decompose it into three meaningful and mutually exclusive components:

- A measure of technical inefficiency arising from the fact that regions are not operating at their benchmark potential.
- A measure of the impact of the quality attributes, which provides at the aggregate level information on the quality/quantity trade-off.
- A measure of the potential for efficiency improvement arising from reallocation of inputs across regions.

The intuition behind this last component is that regions may not be operating on an optimal scale or may be operating with suboptimal input mix (e.g., regions may have shortages of specific inputs); by reallocating resources across regions it is possible to remove these inefficiencies and increase the overall country level output of public service.

The second contribution is a thoughtful discussion on efficiency implications of decentralization/centralization based on the empirical results. In particular, the choice of these two specific countries is determined by interest in the impact of the fiscal structure on the efficiency with which public service is delivered. The UK represents a highly centralized fiscal system, although England in the 18th century is perhaps an example of a wealthy nation under a decentralized regime. Modern China in this paper represents a highly decentralized one, although ancient China has a long history of centralization. Implications in this paper are drawn based on the different sources of inefficiency in each system. Of course, it is recognized that this discussion cannot be directly linked to the empirical results because there may be many other causes for the inefficiency differential observed between China and the UK. Thus, the discussion may be in some way of a speculative nature. Nevertheless, such a discussion is useful in order to frame the results in a more general policy debate related to the efficiency implications of decentralization. The main empirical results point to the fact that in the UK the major source of inefficiency at the aggregate level is due to the reallocation component, while in China the major source of aggregate inefficiency is due to technical inefficiency (lack of best practice). The interpretation proposed in this paper of this result is that decentralized systems (like China) are better at allocating resources efficiently across regions using decentralized information processing, while centralized systems (like the UK) are more capable of pushing the adoption of best practices across regions (at the expense of reallocation inefficiency) with central government overseeing the system and internalizing externalities across regions.

The paper is structured as follows. The next section gives a brief background of the structure of public service provision in each country. The third section describes the methodology used to measure and examine inefficiency in each system. The data and variables used are introduced in the fourth section . Empirical results are presented and discussed in the fifth section. Finally, the sixth section concludes.

THE STRUCTURE OF PUBLIC FINANCE IN MODERN CHINA AND THE UK

Modern China

The evolution of an inter-government fiscal relationship should be understood in the context of economic reform and development following three distinct phases over time: the pre-reform phase prior to 1979, the transitional phase of 1980–93 and the post-1994 phase. The following gives a brief overview of the fiscal reform that has resulted in the current fiscal structure in China. Further details can be found in the literature (e.g. Agarwala, 1992; Jin, Qian, & Weingast, 2005; Shen, Jin, & Zou, 2012).

Prior to the economic reform of 1978–79, the fiscal relations between the central and provincial governments are best described as formed under a consolidated budget system under which the central government set spending priorities (unified spending) and revenues were largely collected from state-owned enterprises in the form of profit and taxes (unified revenue). Local governments lacking discretionary spending power were agents of the central government, just as the state-owned enterprises. Intergovernmental transfers were set to finance the gap between locally collected revenues and permitted local expenditures.

Along with the economic reform, which started in 1978, the central–local government fiscal relations changed significantly. A fiscal revenue sharing system replaced the highly centralized system in 1980 to provide local governments with an incentive to collect revenue. In the period 1988–93, the government implemented a 'fiscal contracting system' under which some provinces had to remit to central government part of their revenues, according to a predetermined lump-sum amount or a progressively increasing ratio of revenues. Central government depended a great deal on this local transfer from the better-off provinces during that period. On average, the local revenue accounted for about 66% of total government budgetary revenue over fiscal contracts.

Starting in 1994, the fiscal contracting system was replaced by a separating tax system. The tax-sharing reform in 1994 explicitly defined fiscal revenue as central revenue, shared revenue and local revenue. Central government changed the revenue-sharing arrangement incrementally after the 1994 reform by adjusting the proportion of the shared revenue that goes to central government upwards.[1] Compared with the previous period, the financing of public service is recentralized, but it is still relatively decentralized. Also, a large portion of expenditure responsibility is devolved to local governments despite the recentralization of tax revenues after the 1994 tax reform.[2] Local governments play the key role in providing social services such as education, healthcare, social security, housing and urban/local services subject to a tightening budget.

In China, a characteristic of fiscal reform is that it goes hand in hand with economic development. Decentralization as a result of the fiscal reform makes it possible for resources allocated for regional public service to be directly linked to regional economic development. Regional economic development determines local government revenue, which then determines how much resources are available for delivering public goods in each region. However, with central government seizing more local government revenue after the 1994 tax reform, this link may not be as tight as before.

The UK

The UK is a country with a relatively high level of fiscal centralization in the sense that local government is part but not a major part of public sector. Local authorities' expenditure accounts for around a quarter of total expenditure of the government and this proportion has changed little for many years (see Figure A2 in Appendix A in the supplemental data online).

Local government expenditure is largely funded by central government. The expenditure by the devolved administrations of the Scottish Executive, Welsh Assembly and Northern Ireland Executive is largely funded by block grant from the UK government. Each devolved administration thus has devolved power in certain service areas subject to the limit of funding from the UK government. In England, local authority expenditure is financed through a balance of a central government grant including non-domestic rates and the locally raised council tax and some other income sources such as capital receipts and investment income. The main services delivered by local government in English regions are primary and secondary education and social services. The main services funded by central government are the National Health Service (NHS), social benefits and pensions, defence, and higher and further education.

Although fiscal power is still relatively centralized at central government, there has been some progress in devolution of power in the past decades. For instance, under the Scotland Act 1998, parliament can pass acts and the Scottish Executive can make secondary legislation in areas other than those reserved to Westminster. In particular, parliament has the power to vary the standard rate of income tax by up to 3 percentage points from the UK level to give an additional source of income. The main service areas in which Scotland has devolved power include health, education, local government, housing, economic development and financial assistance to industry, and some transport.

Under the Government of Wales Act 1998, the assembly can make delegated or secondary legislation, such as orders and regulations, in devolved areas, but primary legislation for Wales in devolved areas is still made by the UK parliament. The National Assembly for Wales has devolved powers mainly in the areas of health, education, local government, housing, economic development, transport and others.

Under the Northern Ireland Act 1998 there is triple division of areas regarding Northern Ireland devolution. The Northern Ireland Assembly can legislate with respect to 'transferred' matters. It can pass both primary and secondary legislation. Other matters are either 'reserved' or 'excepted'. The areas transferred to the assembly mainly include health, education, regional development, enterprise, trade and investment.

In England, the government introduced the Regional Assemblies (Preparations) Bill to parliament in November 2002. These assemblies were to be responsible for regional strategies dealing with sustainable development, economic development, spatial planning, transport, waste, housing, culture (including tourism) and biodiversity. In 2004, a North East referendum for an elected regional assembly took place in which the people of the region delivered a 'no' vote to government. According to University College London (UCL) devolution monitoring reports,[3] the result seemed to be indicative of a more general feeling of disenchantment with politicians. It seems there was also a wider scepticism concerning the capability of devolution, what was perceived to be another layer of bureaucracy and the tax that would accompany it. The future agenda regarding the devolution of English regions is rather unclear. Note that unlike China, in the UK centralization means that resource allocation across regions is largely planned and controlled by central government rather than being directly linked to local fiscal revenue. Studies examining the effect of devolution in the UK are very limited. Pike, Rodriguez-Pose, Tomaney, Torrisi, and Tselios (2012) attempted to quantify 'economic dividend' from UK devolution but concluded that the UK's highly centralized system has marked its particular devolution and established a constrained context for any 'economic dividend' to emerge. They also note the contrast between the UK and more substantive and extensive fiscal decentralization in other countries that have experienced stronger, more widespread, positive effects under certain conditions (also see Rodriguez-Pose & Ezcurra, 2010).

METHODS

The general framework used to measure inefficiency is data envelopment analysis (DEA) using directional distance functions (DDFs). DEA involves the use of linear programming methods to construct a non-parametric piece-wise surface (frontier) of the data. Efficiency measures are then calculated relative to this surface (Charnes, Cooper & Rhodes, 1978; Farrell, 1957).

To construct the surface or frontier, the production possibilities set (or technology, or production set) is described as follows. Consider a region where $\mathbf{x} \in R_+^N$ inputs produce $\mathbf{y} \in R_+^M$ output quantities (public services) with quality attributes $\mathbf{a} \in \mathbb{R}_+^J$. Observations for panel data are collected into three matrices: the input matrix $\mathbf{X}^t = [\mathbf{x}_1^t \dots \mathbf{x}_K^t]$ of dimension $K \times N$ for each time period; the output quantity matrix $\mathbf{Y}^t = [\mathbf{y}_1^t \dots \mathbf{y}_K^t]$ of dimension $K \times M$ for each time period; and the output quality attributes matrix $\mathbf{A}^t = [\mathbf{a}_1^t \dots \mathbf{a}_J^t]$ of dimension $K \times J$. The dataset can be represented by the collection of these matrices:

$$(\mathbf{X}^t, \mathbf{Y}^t, \mathbf{A}^t), \quad t = 1, \dots, T \tag{1}$$

The following assumptions are made in order to define the production possibilities set: (A1) convexity; (A2) free

disposability of inputs; (A3) quality attributes are freely disposable; (A4) output quantities are weakly disposable with respect to quality attributes; and (A5) quality attributes and output quantities are null-joint. The first two assumptions are standard in the DEA approach to efficiency measurement. The third assumption is basically stating that it is possible to dispose freely of quality attributes; in other words, it is possible to produce positive output quantities of no (or poor) quality. A4 implies that disposing of output quantities is only possible by a similar contraction in the overall quality attributes of this quantity. Finally, A5 means that production of zero quantity of output is possible only by producing zero quality. To show how these assumptions are intuitive, this paper reports a simple graphical representation of an output set where inputs are producing one quantity output and one quality attribute. From Figure 1 the following is clear: (1) it is always possible to produce a positive output quantity of no quality; (2) for a given level of quality, output quantity can be reduced only up to a (positive) point; and (3) there is a region showing a trade-off between quantity and quality. Using a recent result of Fare and Grosskopf (2012) these assumptions mean that quantity outputs are jointly limitational for quality attributes, i.e. given a certain quantity of outputs only a finite level of quality can be attained (even by pushing cost/inputs to infinity). Since quantities are limitational for quality, it means that for each level of output quantities there is a maximal attainable level of quality that can be met by expanding cost (or input usage) towards infinity. This is represented in the second panel in Figure 1 with an alternative graphical representation. Here on the x-axis is the input used and on the y-axis the quality of a given output quantity (which is fixed at $\bar{y}$). It is clear that for a given output quantity level, the level of quality that can be attained is limited and is a function of input usage. In other words, for any given level of output quantity, the quality of such output can be increased only by using additional inputs (or additional cost) and there is a limit to the level that can be attained. From a purely technical (and computational) point of view this approach is equivalent to the plant capacity approach proposed by Fare, Grosskopf, and Kokkelenberg (1989). The assumptions made above therefore lead to two different

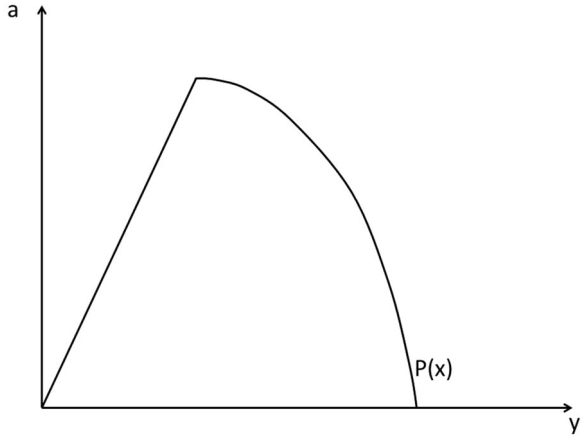

Figure 1. Example of an output set.

benchmarks for a given region being defined: first, an unconditional benchmark that does not take into account the quality attributes; and second, a benchmark conditional on the quality attributes. The idea of the first benchmark is to compare regions that produce the same level of output quantities, irrespective of the quality with which these are produced. Since part of the observed cost differences is a function of the different level of quality attained, in a second stage this is taken into account by separating the effect of quality from the effect of pure inefficiency.

The regional level public service production technology is defined as the variable returns to scale (VRS) envelope of the dataset $(\mathbf{X}^t, \mathbf{Y}^t, \mathbf{A}^t)$, $t = 1, \ldots, T$:

$$\Psi_q^t = \{(\mathbf{x}, \mathbf{y}) : \lambda \mathbf{X}^t \leq \mathbf{x},\ \lambda \mathbf{Y}^t \geq \mathbf{y}^t,\ \lambda \mathbf{A}^t \geq \mathbf{a}^t,\ \sum \lambda_k = 1,\ \lambda \geq 0\} \tag{2}$$

Quality-adjusted technical efficiency (QTE) of public service delivery is measured using the DDF as:

$$\mathrm{QTE} = D^t(\mathbf{x}, \mathbf{y}, \mathbf{a}, \mathbf{g}_x, \mathbf{g}_y) = \sup_\beta \{\beta : (\mathbf{x} - \mathbf{g}_x\beta, \mathbf{y} + \mathbf{g}_y\beta, \mathbf{a}) \in \Psi_q^t\} \tag{3}$$

This definition of inefficiency is looking at possible input quantities reduction and output quantities expansion along the direction given by the fixed numeraire $(\mathbf{g}_x, \mathbf{g}_y)$. The numeraire is interpreted as the unit of measurement of the inputs and outputs; e.g., if input is the number of hours worked in specific regions, then $\mathbf{g}_x$ = 1 hour worked. The regional level potential production technology as the VRS envelope of the reduced dataset $(\mathbf{X}^t, \mathbf{Y}^t)$, $t = 1, \ldots, T$ is defined as:

$$\Psi^t = \left\{(\mathbf{x}, \mathbf{y}) : \lambda \mathbf{X}^t \leq \mathbf{x},\ \lambda \mathbf{Y}^t \geq \mathbf{y}^t,\ \sum \lambda_k = 1,\ \lambda \geq 0\right\} \tag{4}$$

Potential technical inefficiency (TE) of public service delivery is defined as:

$$\mathrm{TE} = D^t(\mathbf{x}, \mathbf{y}, \mathbf{g}_x, \mathbf{g}_y) = \sup_\beta \{\beta : (\mathbf{x} - \mathbf{g}_x\beta, \mathbf{y} + \mathbf{g}_y\beta) \in \Psi^t\} \tag{5}$$

The second definition eliminates the constraint associated with output qualities and looks at the potential input saving and output expansion that can be achieved. Another way of interpreting this quantity is the following: suppose a quality benchmark is ignored and one is ready to give up all the quality in order to reach a higher output quantity target; then the quantity embedded in TE is giving this potential expansion. The above discussion about limitationality means that TE $\geq$ QTE and that the impact of the quality attributes on overall input usage will be measured as the difference between these two measures (this is also true from a technical point of view because some of the constraints are being omitted in order to compute TE):

$$\mathrm{ATE} = \mathrm{TE} - \mathrm{QTE} \tag{6}$$

TE is higher than QTE by construction based on what is considered to be a reasonable assumption that for a given

cost there is a choice to be made (trade-off) between quality and quantity.[4] In other words, it is not possible to increase quality without either reducing quantity or increasing cost. Therefore, ATE (which is the difference between TE and QTE in equation (6), thus the impact that observed quality makes) being greater than zero reflects the assumption that the impact of achieving the observed level of quality is to trade-off (not to enhance) some of the potential output expansion given input. Even though by construction TE > QTE (therefore ATE > 0), the magnitude of ATE still very much depends on empirical data.

Therefore, this indicator just defined in equation (6) has a neat interpretation: it is the amount of input–output quantity that the regions must give up in order to attain the observed level of quality. In other words, it can capture the trade-off between quality and quantity by means of a single number for each individual region. Since the trade-off between quantity and quality is formally defined, the next task is to proceed to the definition of aggregate measures of performance at the country level.

The group or aggregate level production possibilities set is given as the sum of S identical regional production possibilities sets (Fare, Grosskopf, & Zelenyuk, 2008; Nesterenko & Zelenyuk, 2007; Li & Ng, 1995):

$$\Psi^t(S) = \sum_{j=1}^{S} \Psi^t \tag{7}$$

The country-level production possibility set is defined as the union of all possible aggregate production technologies (Peyrache, 2013):

$$\Psi_I^t = \bigcup_{S=1}^{+\infty} \Psi^t(S) \tag{8}$$

Figure 2 shows these different technology sets. Country-level potential technical inefficiency (IE) is defined as the potential expansion of output quantities and contraction of inputs when the country-level technology is used as a benchmark:

$$\begin{aligned} \mathrm{IE} &= D_I^t(\mathbf{x}, \mathbf{y}, \mathbf{g}_x, \mathbf{g}_y) \\ &= \sup_\beta \left\{ \beta : (\mathbf{x} - \mathbf{g}_x\beta, \mathbf{y} + \mathbf{g}_y\beta) \in \Psi_I^t \right\} \end{aligned} \tag{9}$$

This optimization problem can also be represented (in an equivalent way) by the following integer linear programme:

$$\begin{aligned} \Psi^t &= \sup_{\lambda, K^*} \beta \\ \text{s.t. } \lambda \mathbf{X}^t &\le \mathbf{x} - \mathbf{g}_x\beta \\ \lambda \mathbf{Y}^t &\ge \mathbf{y} + \mathbf{g}_y\beta \\ \sum \lambda_k &= K^* \\ \lambda &\ge 0 \end{aligned} \tag{10}$$

where K^* is an integer to be determined by the optimal solution. The optimal value of the intensity constraint in the previous definition is interpreted as the optimal number of regions that should populate the country in order for public service to be delivered efficiently. The efficiency indicator at country level gives a measure of total input waste and output loss at the country level as a whole. It is important to emphasize that definitions (5) and (9) differ because of the different benchmark technology used: definition (5) uses regional-level production technology while definition (9) uses country-level production technology. Since all the differences come down to these two definitions of technology, Figure 2 shows the two different technologies used to compute the two alternative indicators.

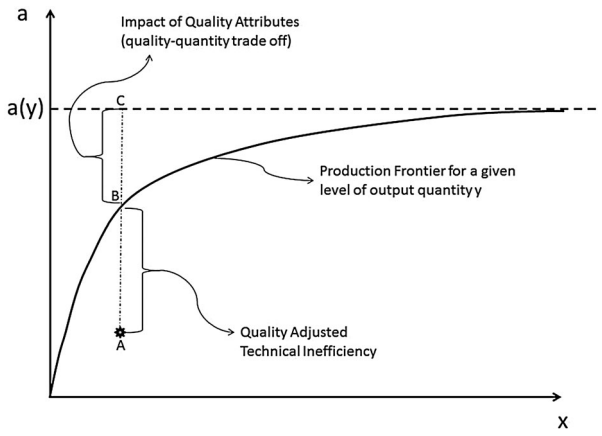

Figure 2. Technical efficiency decomposition.

Since DDF is an absolute measure of inefficiency expressed in a given common numeraire for all the regions, it is possible to compute it for each region in the dataset and sum it up into an index of country technical inefficiency:

$$\mathrm{ITE}^t = \sum_{k=1}^{K} D^t(\mathbf{x}_\mathbf{k}^\mathbf{t}, \mathbf{y}_\mathbf{k}^\mathbf{t}, \mathbf{g}_x, \mathbf{g}_y) \tag{11}$$

This indicator is a measure of waste in inputs and loss in outputs at the country level, due to the technical inefficiencies of the regions actually operating in the country. The total observed inputs and outputs at time t for the entire country is then defined as:

$$\mathbf{I}^t = \sum \mathbf{x}_k^t, \mathbf{Q}^t = \sum \mathbf{y}_k^t$$

A measure of country inefficiency is given by the following mixed-integer linear programme:

$$\mathrm{IE}^t = D_I^t(\mathbf{I}^t, \mathbf{Q}^\mathbf{t}, \mathbf{g}_x, \mathbf{g}_y) \tag{12}$$

Even if all the regions in the country are technical efficient (i.e., ITE = 0), the country as a whole could still be inefficiently organized (i.e., $\mathrm{IE}^t > 0$). The discrepancy between the two indicators is a measure of reallocation inefficiency at the country level:

$$\mathrm{RE}^t = \mathrm{IE}^t - \mathrm{ITE}^t \tag{13}$$

This indicator is always larger than zero ($\mathrm{ITE}^t \le \mathrm{IE}^t$) and represents the inefficiency arising from the way public service delivery is structured across regions. This discrepancy

is attributable to input reallocation across regions and it arises because of possible scale and input mix effects. Now that the quality-quantity trade-off is represented via an overall indicator based on DDF and that the indicator is computed for each region, it is possible to sum up all these indicators into an aggregate measure of quality attributes:

$$\mathrm{IQTE}^t = \sum_{k=1}^{K} \mathrm{QTE} \tag{14}$$

The country-level technical inefficiency measure is therefore decomposed into:

$$\mathrm{ITE}^t = \mathrm{IQTE}^t + \mathrm{IATE}^t \tag{15}$$

Therefore, inserting this expression into the country-level decomposition, a structural decomposition of country-level inefficiency into the following three components is obtained:

$$\mathrm{IE}^t = \mathrm{IQTE}^t + \mathrm{IATE}^t + \mathrm{RE}^t \tag{16}$$

The first component measures how much input is wasted and output is lost because of technical inefficiency of individual regions. The second component is an aggregate measure of the quality–quantity trade-off: it tells how much additional input–output quantity is used in order to secure the observed level of output quality. The last component measures potential efficiency improvements that may arise from the reallocation of inputs across regions. A more direct interpretation of this decomposition can be provided in percentage terms:

$$\%\mathrm{IQTE}^t + \%\mathrm{IATE}^t + \%\mathrm{RE}^t = 1 \tag{17}$$

where:

$$\%\mathrm{QTE}^t = \frac{\mathrm{IQTE}^t}{\mathrm{IE}^t}$$

$$\%\mathrm{IATE}^t = \frac{\mathrm{IATE}^t}{\mathrm{IE}^t}$$

$$\%\ \mathrm{RE}^t = \frac{\mathrm{RE}^t}{\mathrm{IE}^t}$$

This transformation gives a more direct interpretation of how much the different components contribute to the overall inefficiency of the country.

DATA AND VARIABLES

Public expenditure in the UK is planned and controlled at the central level on a departmental basis. For the country and regional analysis, total managed expenditure is divided into identifiable and non-identifiable expenditure. Identifiable expenditure can be recognized as having been incurred for the benefit of individuals, enterprises or communities within particular regions. Examples are mostly health, education and transport services, and spending on social security and on pensions. It also includes local collective services which are consumed locally such as all local authority spending; central government spending on regional development agencies; central government spending on police and local courts. Non-identifiable expenditure is the one that cannot be so identified, for example because it is deemed to be incurred on behalf of the UK as a whole, e.g. defence expenditure, overseas representation, tax collection and some environmental protection spending. This paper uses identifiable regional expenditure (real) at the regional level as input. It covers the service areas of education and health. The selection of this coverage is to ensure that the data are as consistent as possible over time and across both the UK and China. Although the data of UK and China are not pooled, the authors made sure all the variables described below were reasonably consistent across the two countries for the analysis and discussion of the results.

In contrast to the top-down approach in the UK, China adopts the bottom-up approach for public expenditure. In fact, the China Statistic Yearbooks only provide statistics on local government expenditure at a regional level but not central government expenditure allocated at a regional level. However, after the tax reform in 1994, the budgeted local government expenditure is financed by local tax revenue, shared tax revenue and tax refund (transfer) from central government. Therefore, the reported local government expenditure is partly financed by central government.

Before 2007 there was no international standard in China for the budget classification – the use of the International Monetary Fund's (IMF) *Government Financial Statistics Manual* (GFS) (2001) for the economic classification and the United Nation's *Classification of Functions of Government* (COFOG) for the functional classification. The Ministry of Finance in China implemented the GFS 2001 in 2007. This now forms the basis for governments at different levels to budget and report fiscal revenue and expenditure statistics.[5] Therefore, the classifications of government expenditures at regional level explained below have changed since 2007. But this should not have too big an impact on the results as the aggregate expenditure was used as the single input. Consistent with the UK data, government real expenditure at the regional level, which covers the main service areas of education and public health, is used. The data show that the subtotal expenditure covering selected public service areas exhibits similar patterns to the total expenditure across the regions in both countries (Figures A3 and A4 in Appendix A in the supplemental data online show the pattern in local government expenditure across the regions and over the years in the UK and China).

Note that the use of real expenditure as the only input in the model will generate efficiency scores that incorporate possible effects of inputs price differences across regions. For example, if it is cheaper to hire a teacher in Scotland than it is in London, then for the same level of education output London will result in a higher expenditure compared with Scotland, since the expenditure variable will include a price effect. This will be reflected in the efficiency scores. Unfortunately, one encounters here a limitation in the data because it is very difficult to obtain data on input usage and input prices at the regional level. On a

more positive note, this effect of prices may be attenuated greatly in public service. In fact, one may expect the difference in salaries and the price of other inputs in the public sector not to vary much across regions. However, to mitigate this problem at least to some extent, the expenditure is de-trended, as explained below.

Regarding outputs, the following indicators are used to reflect the outcome of public service delivery in health and education.

In terms of education, the total number of pupils on roll in both primary and secondary schools (this refers to the public funded school sector in the UK) is used as the output quantity measure to indicate the coverage of the public service of education. In addition, two variables are used to indicate the quality of the education service. The first is the number of teachers; it is used as an indicator of teachers' availability in both China and the UK. The number of teachers is considered as a quality measure because given the size of the student cohort (education coverage) increasing the number of teachers will increase the quality of education.[6]

The second quality variable aims at capturing student achievement. In the UK, the General Certificate of Secondary Education (GCSE) (General National Vocational Qualification (GNVQ) in Scotland) achievement measured as the percentage of 15-year-old pupils achieving five or more GCSEs at grades A*–C is used to measure student achievements. As for China, such a variable is not available. Instead, the number of university students is used to indicate the achievement of secondary school students. Entrance to universities is highly competitive in China and conditional on successful secondary education attainment. In every region students have to achieve a minimum level of total scores in the entrance examinations to obtain entry into universities and the number of university students will then depend on the number of secondary students achieving this threshold. The authors therefore believe that in such a competitive and score-based university entrance system, the number of university students should be highly associated with secondary school students' achievement. The competitive nature of Chinese higher education admission is also consistent with the World Bank World Development Indicators (WDIs): the enrolment rate of tertiary education (International Standard Classification of Education (ISCED) 5 and 6) in China is 23% (of the population of the five-year age group following on from secondary school leaving) in 2010 in contrast with 61% in the UK in the same year. Therefore, without access to data of direct measurement of secondary school students' achievement or the number of secondary school graduates who attend universities,[7] one has to consider that the number of higher education students given the size of population in the region is a good proxy for secondary school students' achievement. Moreover, this variable should be able to capture differences in the quality of education across Chinese regions, considering that the entrance tests set by the universities are standardized across the nation.

In the area of public health the authors start with the consideration that healthcare is produced to cover the whole population in the region. Thus, population size is used as a quantity measure. Two quality variables are as follows: number of licensed doctors and number of hospital beds (for each region). These two measures are considered as good proxies for quality because given the coverage of public health care (size of population) increasing the number of doctors and hospital beds should increase the quality of healthcare.[8]

A variable measuring the total area of each region in each country is also introduced. This choice is dictated by the fact that public service delivery may be subject to economies of density. This means that other things being equal (i.e. same population, same number of students, same quality outcome), in a region with a larger area it will cost more to deliver the same public service. So, for example, it will be cheaper to provide healthcare in Shanghai (where the population is large relative to the area) than in Tibet (where the population is small relative to the area). The total area of each region is treated as a quantity output in the model, mainly for the reason that, *ceteris paribus*, increasing the area will either increase total cost or decrease some of the other output quantities.

The complete data sample as a result of the collection of the above variables is 12 regions (nine English office regions and Scotland, Wales and Northern Ireland) over the period 2000–10 in the UK (in total 372 observations) and 31 provinces/regions from 1999 to 2010 in China (in total 132 observations).[9] Considering the potential large difference in terms of institutional environment across the two countries, this paper analyzes the efficiency of public service using the two samples separately rather than pooling them.[10] This is, of course, suboptimal since it would be interesting to have datasets that are comparable across the two countries. Unfortunately, here the paper hits a limitation of the data, since the statistical agencies of the two countries are not harmonized and comparable data do not exist at this point (to the best of the authors' knowledge). Though the two datasets are not pooled, a comparison can still be made in terms of efficiency components because efficiency is a benchmark notion (i.e. each region in each country is compared with the best region in that country). Tables A1 and A2 in Appendix A in the supplemental data online show the summary statistics of the variables and their sources. Box plots for all variables in both countries are reported in Figures A5 and A6, also in Appendix A.

In both datasets there was quite a significant proportion of increase in real expenditure over time that was not explained by the output variables (this may come from inaccuracies of the deflator index for the expenditure variable or some input prices which are increasing in time). In order to obtain meaningful results in the DEA analysis (which is run on pooled data for each country separately[11]) the cost series are de-trended by using a standard regression approach. A cost function is estimated by regressing observed cost onto the output variables and including a linear time trend. A Cobb–Douglas and a translog

Table 1. China decomposition (output-oriented model).

Year	%(ITE/IE)	%(IQTE/ITE)	%(IATE/ITE)	%(RE/IE)
1999	64.3	22.9	41.4	35.7
2000	54.8	24.3	30.6	45.2
2001	50.2	20.5	29.7	49.8
2002	48.9	18.5	30.3	51.1
2003	54.2	22.9	31.3	45.8
2004	62.8	19.2	43.6	37.2
2005	53.8	19.7	34.1	46.2
2006	56.7	27.4	29.3	43.3
2007	60.5	22.1	38.4	39.5
2008	65.0	25.1	39.9	35.0
2009	73.4	32.5	40.9	26.6
2010	62.6	19.7	42.9	37.4
Mean	58.9	22.9	36.0	41.1

specification were used for the cost functional form. Both ordinary least squares (OLS) and fixed effects are used with both functional specifications. In all cases the data fit is very good and the coefficient of the time trend is significant. The results indicate a trend of around 15% for China and 5% for the UK (this is the part of cost increase that cannot be explained by outputs). Since the fixed effects translog model is the most general (it nests all the others) and since the standard errors of this model are not much higher than the other models, the time trend coefficients associated with this model are used to de-trend the cost series (it should also be emphasized that the other models may suffer from misspecification and inconsistency). The fact that all models point to more or less the same value lead the authors to believe that there was an increase in cost dictated by some type of inflation that was not well captured by the gross domestic product (GDP) deflator. The de-trended cost series were calculated applying the rates of increase implied by the fixed effects translog model, i.e. 5% for UK and 15% for China.

Table 2. UK decomposition (output-oriented model).

Year	%(ITE/IE)	%(IQTE/ITE)	%(IATE/ITE)	%(RE/IE)
2000	36.6	11.1	25.5	63.4
2001	34.8	8.9	25.9	65.2
2002	36.0	14.4	21.6	64.0
2003	37.5	13.1	24.4	62.5
2004	39.3	13.6	25.7	60.7
2005	41.8	15.0	26.8	58.2
2006	45.9	16.1	29.8	54.1
2007	44.3	15.6	28.7	55.7
2008	46.1	15.4	30.7	53.9
2009	47.5	5.0	42.5	52.5
2010	44.4	1.4	42.9	55.6
Mean	41.3	11.8	29.5	58.7

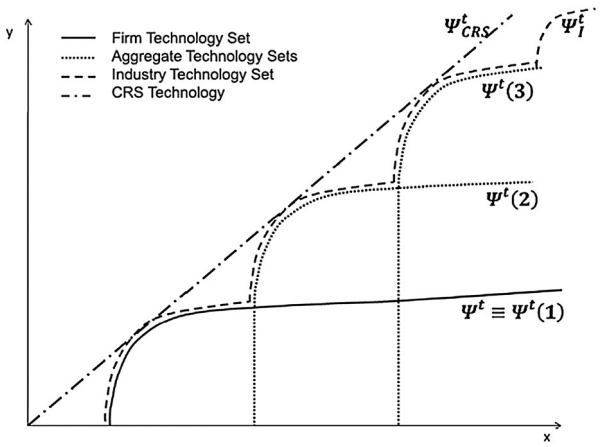

Figure 3. Representation of different technologies for the one input–one output case.

MAIN RESULTS AND DISCUSSION

Tables 1 and 2 report the DEA decomposition analysis with the values expressed in percentage terms (the same decomposition expressed in absolute values is reported in Tables A3 and A4 in Appendix A in the supplemental data online). In the empirical analysis the following directional vectors were used:

$$g_x = 0, g_y = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \text{ and } g_z = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}$$

The directional vectors just specified consider an expansion of output quantities in health and education, while keeping all the other variables constant: area, cost and all quality measures. In other words, the question addressed is how much population growth and increase in the number of student enrolments can be met without increasing expenditure or reducing the quality of the service. This choice is derived from interest in knowing by how much public service delivery could be expanded.[12] By looking at the percentages the differences are striking. The most important component in China is technical inefficiency (58.9%), while in the UK this component accounts only for about 41.3%. Note that technical inefficiency arises from individual regions falling short of best practices. On the contrary the reallocation component accounts for only about 41.1% in China and for 58.7% in the UK. As mentioned in the methodology section, reallocation inefficiency is attributable to input reallocation across regions and it arises because of possible scale and input mix effects. Finally, the impact of quality shows different trends too, accounting for more than 70% of UK overall technical inefficiency and about 60% of China inefficiency. In other words, the impact of quality is stronger in UK regions than in China regions: a larger part of the loss in output expansion in UK regions when compared with regions in China is attributed to the observed level of quality differences.

The rationale behind the choice of the two countries in this paper was that China is an example of a fiscally decentralized system, while the UK is a highly centralized fiscal

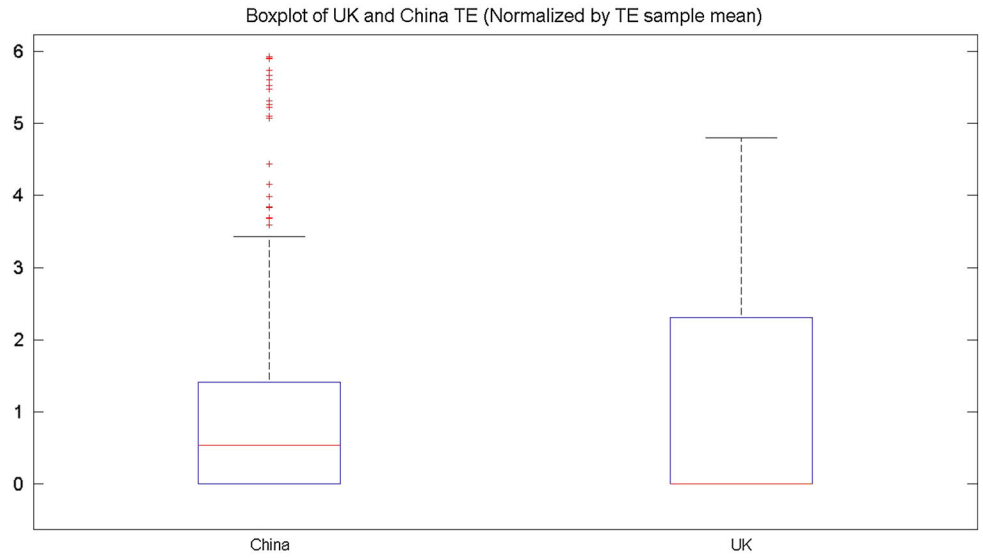

Figure 4. Box plot of inefficiency scores (normalized by sample mean).

system. By examining the results reported in the previous tables, it is quite clear that in the decentralized system (China) allocation of resources for public service delivery (the reallocation component) seems to be well addressed. This is mirrored by a highly disperse level of ITE across regions, which points to the possibility that best practices are not effectively enforced via decentralization. On the contrary, the centralized system (UK) is quite effective in pushing the adoption of best practices, as indicated by the low weight ITE has on overall country inefficiency; this comes at the expense of a high level of reallocation inefficiency: the UK (unlike China) seems to be less able to allocate resources efficiently across regions. Figure 3 shows clearly the greater dispersion of overall efficiency scores (IE) in China compared with the UK.

Can one make sense of this result?

The differences in the relative weight of the different components for the two countries are remarkable. Even though this paper does not formally test if these differences are caused by the way the fiscal structure is organized, it is possible to make an argument that the effect of decentralization on public service delivery efficiency may not be neutral. It seems to indicate a sort of trade off between efficient resource allocation across regions and efficient best-practice operation at the regional level. Such a trade-off is quite consistent with various strands of literature related to the theories of firms and theories of public finance.

In principle, if a complete contract is possible, then central government can design it in order to obtain information from local government (see Myerson, 1982, for a statement of the revelation principle) and use it to allocate resources efficiently. Incompleteness of contracts means that, under decentralization, local government has greater incentives to obtain local information than central government (Cremer, Estache, & Seabright, 1994). In addition, even if central and local governments have symmetric information, theories of organization suggest that decentralized information processing under the setting of constrained information processing capacity of any single centralized planner could spontaneously coordinate more efficient resource allocation (for a review of decentralized information process and theories of organization, see van Zandt, 1996; & van Zandt, 1997). With local government's information advantage (also see Oates, 1972; and Tiebout, 1956, for the traditional theory of public finance arguing for decentralization) and decentralized information processing, it is therefore not surprising that decentralization deals with allocation of inputs across regions more efficiently than centralization.

Local governments may also have the incentive to use information for their own interests, and this incentive costs may grow with increasing vertical layers (see Mookherjee, 2006, for a review of modelling incentives in mechanism design theory). Since centralization means there is effectively one vertical layer only, incentive costs may be more serious with decentralization. This could perhaps explain why individual regions in a decentralized system such as China have more disperse technical inefficiency than the UK. Moreover, similar to the incompleteness of contracts between local and central government regarding local information, contracts between regions will be incomplete and consequently there will be limits to the extent to which British externalities can be British-internalized purely by bargaining between regions under American decentralization (Seabright, 1996). Centralization by allowing central government to control and oversee the overall system could internalize externalities between regions,[13] and this could perhaps reduce heterogeneity in performance across regions and push individual regions to achieve their potential benchmark (as observe here with the UK regions).

CONCLUSIONS

This paper examines the sources of inefficiency of public service delivery in two countries, the UK and China. The

proposed methodology allows one to examine the two countries separately (using different datasets with homogeneous variable definitions), while still being able to provide some degree of cross-country comparison. This is achieved by computing the inefficiency at country level based on regional efficiency estimates. Country-level inefficiency is then decomposed into three components that indicate different sources of inefficiency. This decomposition allows one to account explicitly for quality attributes of public service delivery via an indicator that completely quantifies the trade-off between output quantity expansion (given input) and observed quality outcomes. The differences in terms of sources of inefficiency in the two countries are remarkable. China shows a relatively efficient allocation of resources across regions in contrast to a highly dispersed level of regional technical inefficiency. On the contrary, in the UK technical inefficiency at a regional level is relatively low compared with the high level of inefficiency of resource allocation across regions. This paper argues this could indicate a potential trade-off between efficient resource allocation across regions and efficient delivery of public service at a regional level under different fiscal structures. This may be rationalized using existing off-the-shelf theoretical literature. Decentralized systems, while promoting a more efficient allocation of resources with decentralized information processing capacity, are more inclined to suffer from distorted incentive costs (which increase with the number of vertical layers), which may lead to a lack of best practices. On the contrary, a centralized system may be more capable of facilitating the adoption of best practices by overseeing the whole system, but at the cost of generating an inefficient allocation of resources with constrained information processing capacity of central planners.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors

FUNDING

The financial support of the Economic and Social Research Council (ESRC) is gratefully acknowledged by Minyan Zhu.

SUPPLEMENTAL DATA

Supplemental data for this article can be accessed at http://10.1080/00343404.2015.1080992

NOTES

1. For instance, from May 1997 the sharing ratio of stamp taxes on security exchange between the central and local governments was adjusted from 50%/50% to 88%/12%; from 1 October 2000 it was changed to 97%/3% in the subsequent three years; since 1 January 2002 the central and local governments share all the company income tax revenues, except a list of enterprises, and personal income tax revenues together at the ratio of 50%/50% in 2002. In 2003 and 2004 the central government's sharing rate went up to 60%.
2. Figure A1 in Appendix A in the supplemental data online shows the change of local government revenue and expenditure over the years. There is a significant drop in the ratio of local government revenue after 1994, while the ratio of local government expenditure increases steadily over time. The reported statistics by the Chinese government are misleading in the sense that there is a large gap between reported local revenue and local expenditure in Figure A1. This is because the reported revenue includes local tax revenue and shared taxed revenue, but not transfers from central government (such as tax rebates and equalization grant).
3. See http://www.ucl.ac.uk/constitution-unit/research/research-archive/archive-projects/devolution-monitoring 99-05/.
4. This assumption is not that different from the choice of relevant inputs and outputs to be included in the model. When deciding that a variable is an output, a statement is being made about the effect that an increase of this variable has on the efficiency score.
5. See Communication from the Ministry of Finance, 28 February 2011.
6. It is understood that the number of teachers provided can be considered as an input measure in the situation where the emphasis is on student learning. The number of teachers is considered as an outcome measure here since the number of teachers provided at school level is associated with government expenditure – which is an input variable. Of more concern here is the efficiency of the delivery of public service by government. In this context, this indicator is considered to be important as it is closely related to the amount of money spent on school children. However, it is recognized that the link between pupil–teacher ratio and education achievement may not be clear-cut and also depend on other factors such as class size.
7. It is not recognized that the number of university students is also a reflection of the supply of university places and other factors such as education policies and family income levels. This measure could be improved by future data on secondary school examination scores or similar measures that directly assess secondary education achievement in China. But as long as the supply of university places is not completely rigid, the measure should capture the element of secondary school performance to some extent.
8. A related issue has to do with the somewhat intangible nature of the quality of public services. In principle it would be interesting to include some measure of citizen satisfaction that could proxy for such intangible variables. Unfortunately, at this stage no suitable variable that could accomplish this task was found.
9. In principle, it could be possible to use a finer distinction than just macro-regions. The choice was dictated by keeping the analysis at some reasonable level of aggregation, while making the distinction between 'regional' and

'national' possible. For example, in both these countries some of the decision-making associated with health and education may be made at a lower level of government (e.g. city level instead of province level). Nevertheless, no matter at which level the decision is made about how to use resources, the question of whether these resources are used efficiently can be posed at a more meso–macro-level. In fact, inefficiency that can arise from decision-making at a micro-level will be reflected at the macro-level.

10. Data for each country have been pooled, i.e. all the regions in the UK were pooled and all the regions in China were pooled, but the UK and China were not pooled.

11. The DEA provides upward-biased efficiency scores (i.e. downward-biased estimates of the frontier). The bias vanishes asymptotically; therefore, different sample sizes will return different biases.

12. In the analysis using the directional distance function, the choice of direction inevitably involves some degree of value judgement by the researcher. See Peyrache and Daraio (2012) and Daraio and Simar (2014) for discussions of the effects of direction choice on technical efficiency scores.

13. This benefit comes at the expense of reduced government accountability.

REFERENCES

Agarwala, R. (1992). *China: Reforming intergovernmental fiscal relations*. (Discussion Papers No. 178). Washington, DC: World Bank.

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

Cremer, J., Estache, A., & Seabright, P. (1994). *The decentralization of public Services: Lessons from the theory of the firm*. (Research Working Paper No. WPS 1345). Washington, DC: World Bank. Retrieved from http://documents.worldbank.org/curated/en/1994/08/698107/decentralization-public-services-lessons-theory-firm.

Daraio, C., & Simar, L. (2014). *Efficiency and benchmarking with directional distances: A data driven approach*. (Technical Report No. 7). Rome: Department of Computer, Control and Management Engineering, Università degli Studi di Roma 'La Sapienza'.

Fare, R., & Grosskopf, S. (2012). *Regulation and unintended consequences*. (CERE Working Paper No. 17). Umea: Centre for Environmental and Resource Economics.

Fare, R., Grosskopf, S., & Kokkelenberg, E. C. (1989). Measuring plant capacity, utilization and technical change: A nonparametric approach. *International Economic Review*, *30*, 655–666. doi:10.2307/2526781

Fare, R., Grosskopf, S., & Zelenyuk, V. (2008). Aggregation of Nerlovian profit indicator. *Applied Economics Letters*, *15*, 845–847.

Farrell, M. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society Series A*, *120*, 253–281. doi:10.2307/2343100

International Monetary Fund (IMF) (2001). *Government finance statistics manual*. Washington, DC: IMF. Retrieved October 27, 2015 from https://www.imf.org/external/pubs/ft/gfs/manual/

Jin, H., Qian, Y., & Weingast B. R. (2005). Regional decentralization and fiscal incentives: Federalism, Chinese style. *Journal of Public Economics*, *89*, 1719–1742. doi:10.1016/j.jpubeco.2004.11.008

Li, S., & Ng, Y. (1995). Measuring the productive efficiency of a group of firms. *International Advances in Economic Research*, *1*, 377–390.

Mookherjee, D. (2006). Decentralization, hierarchies, and incentives. *Journal of Economic Literature*, *44*, 367–390. doi:10.1257/jel.44.2.367

Myerson, R. B. (1982). Optimal coordination mechanisms in generalized principal agent problems. *Journal of Mathematical Economics*, *10*, 67–81. doi:10.1016/0304-4068(82)90006-4

Nesterenko V., & Zelenyuk, V. (2007). Measuring potential gains from reallocation of resources. *Journal of Productivity Analysis 28*, 107–116.

Oates, W. (1972). *Fiscal federalism*. New York: Harcourt, Brace Jovanovich.

Peyrache, A. (2013). Industry structural inefficiency and potential gains from mergers and break-ups: A comprehensive approach. *European Journal of Operational Research*, *230*, 422–430. doi:10.1016/j.ejor.2013.04.034

Peyrache, A., & Daraio, C. (2012). Empirical tools to assess the sensitivity of directional distance functions to direction selection. *Applied Economics*, *44*, 933–943. doi:10.1080/00036846.2010.526582

Pike, A., Rodriguez-Pose, A., Tomaney, J., Torrisi, G., & Tselios V. (2012). In search of the 'economic dividend' of devolution: Spatial disparities, spatial economic policy, and decentralisation in the UK. *Environment and Planning C: Government and Policy*, *30*, 429–444. doi:10.1068/c10214r

Rodriguez-Pose, A., & Ezcurra, R. (2010). Does decentralization matter for regional disparities? A cross-country analysis. *Journal of Economic Geography*, *10*, 619–644. doi:10.1093/jeg/lbp049

Seabright, P. (1996). Accountability and decentralisation in government: an incomplete contracts model. *European Economic Review*, *40*, 61–89. doi:10.1016/0014-2921(95)00055-0

Shen, A. C., Jin J., & Zou H.-F. (2012). Regional decentralization and fiscal incentives: Federalism, Chinese style. *Annals of Economics and Finance*, *13*, 1–51.

Smith, A. (1776/1904). *An inquiry into the nature and causes of the wealth of nations* [1776], edited by Edward Cannen, 5th Edn. London: Methuen & Co.

Tiebout, C. (1956). A pure theory of local expenditures. *Journal of Political Economy*, *64*, 416–424. doi:10.1086/257839

Van Zandt, T. (1996). Decentralized information processing in the theory of organizations. In Sertel M. (Ed.), *Contemporary economic development reviewed, Vol. 4: The enterprise and its environment* (pp. 125–160). London: Macmillan.

Van Zandt T. (1997). *Real-time hierarchical resource allocation* Retrieved October 27, 2015 from https://www.kellogg.northwestern.edu/research/math/papers/1231.pdf.