Vol. 13: pp1, 2013
doi: 10.3354/esep00132

ETHICS IN SCIENCE AND ENVIRONMENTAL POLITICS
Ethics Sci Environ Polit

Published online May 2

Contribution to the Theme Section 'Global university rankings uncovered'

FREE ACCESS

World class universities and international rankings

David A. Turner

Faculty of Business and Society, University of South Wales, Pontypridd, CF37 1DL, UK

ABSTRACT: This paper examines 2 of the major international university rankings, the Shanghai Jiao Tong University ranking and the *Times Higher Education Supplement* ranking, and asks why, when they use such different measures of performance, they produce such similar rankings. The author introduces Data Envelopment Analysis (DEA) and shows how this can be an effective method for setting up a range of ranking systems. This is used to explore a number of key issues that are raised by league tables, including which indicators to select, and whether to treat them as inputs or outputs. The paper then examines the assumptions that would need to be incorporated into the DEA approach in order to produce a league table that is similar to the 2 extant examples. It concludes that the currently accepted league tables assume that costs are of no interest, and that high quality outputs are to be valued at any cost. This raises the question of whether such rankings are appropriate for an age of austerity, or whether value-for-money rankings might not be more appropriate.

KEY WORDS: University · Ranking · League tables · World class · Data Envelopment Analysis · DEA

INTRODUCTION

University league tables and rankings are now serious business, forming the basis of national policy for higher education in some cases, and attracting the attention of the World Bank. Salmi (2009) suggests that there are policy lessons to be taken from the league tables at both national and institutional level. In a somewhat lighter vein, Birnbaum (2007) suggested the 'Borges System' for coping with the proliferation of ranking systems in current use.

Birnbaum refers to a short story by Borges (1970) entitled 'The Library of Babel'. The eponymous library included every possible book that could be printed from a fixed alphabet, where the letters had been assigned at random and in random order. The outcome was a huge collection of books, most of which contained printed rows of random letters that made no sense whatsoever. The job of scholars in the library was to go from shelf to shelf looking for books that contained occasional snippets of language that actually made sense. Borges has defined his universe, by specifying the number of pages in each book, the number of lines on each page, the number of characters on each line and the number of characters (letters, punctuation marks and space) that can be printed in each space. It is then a relatively simple matter to calculate the permutations possible to work out the number of possible books in the library. By analogy, Birnbaum invites us to take the whole range of possible university rankings, and select from those the ones that include a fragment of sense.

That is the procedure that I intend to follow in this paper. I will first examine the range of possible ranking systems. I will then examine the ranking systems that are actually current in the world, and see what it is that those systems have in common. The main contenders for consideration as definitive ways of ranking institutions are the systems of the *Times Higher* – Quacquarelli Symonds (QS) (2008) and the Shanghai Jiao Tong University (2008) rankings (now the Academic Ranking of World Universities). Lastly, I shall address the question of whether any of the ranking systems, possible or actual, contain any element of

*Email: david.turner@southwales.ac.uk

sense. However, since the range of possible ranking systems is not as well defined as the individual volumes in Borges' library, perhaps we should start exercising some imagination as to what the realm of all possible world ranking systems of universities could look like.

CHOICE OF INDICATORS

In constructing a ranking system, a number of indicators must be selected in order to form the basis for calculating the ranking. Salmi (2009) distinguishes between 'subjective' indicators and 'objective' indicators. He suggests for example, that the *Times Higher* – QS uses more subjective measures, as it uses indicators of esteem, and the reputation of institutions as reported by academics and employers. In contrast, the Shanghai Jiao Tong University rankings use more objective measures, such as the number of Nobel laureates in an institution.

However, it can fairly readily be seen that the distinction is a false one if we try to consider what the possible universe of ranking systems could be. We might pick a useful indicator of research quality, such as the average number of bicycles on campus during the daytime. (If anybody should doubt that is a measure of quality, they might consider where Harvard, Oxford and Cambridge would appear in a ranking system that used such an indicator.) This is clearly an 'objective' indicator, in the sense that a number of bicycles is quite definite, and, at least in principle, countable. However, as will immediately be apparent, its selection as an indicator of research quality is subjective, which is to say, the compiler of the ranking thinks the indicator appropriate.

If we take, for example, the number of Nobel laureates and Fields Medal winners, an indicator used by the Jiao Tong rankings, there have only ever been just over 800 Nobel Prizes awarded, 20 to organisations and 789 to individuals (Nobel Foundation 2009). Fewer than 50 Fields Medals have been awarded. It follows that this measure of (supposed) research quality can depend upon only a handful of people at even the strongest of institutions. It is certainly not representative of the rank and file activity of members of the university, as, for example, a census of bicycle use would be. But it raises questions about what the number of Nobel laureates in an institution actually indicates. What are the circumstances that attract these scholars to an institution? Are we looking at anything more than another indicator of reputation? There is nothing wrong with using the number of Nobel laureates as an indicator of the research quality of an institution, so long as one recognises that it represents the work of a tiny minority of scholars in a relatively limited range of subject areas (with the emphasis on the physical sciences and medicine). It only becomes dangerous if it is taken as a measure of something else, and is presumed to have policy implications at the national and institutional level.

The arbitrary nature of the selection of indicators is further illustrated by the fact that the Jiao Tong rankings only take into account Nobel Prizes in Physics, Chemistry, Medicine and Economics, thus excluding the Nobel Prizes for Peace and Literature. However good the rationale for such a choice, it clearly implies a particular notion of what is to be counted as 'research', a notion that might not coincide with the activities of all universities around the world.

There is a common belief, based again on the natural sciences, that in order to achieve research of high quality a 'critical mass' of scholars in a single institution is required. This is repeated by Salmi (2009) and is used as the basis for policy recommendations. However, a handful of people in an institution that embraces upwards of 10 000 people can only be called a 'critical mass' by an extreme stretch of the imagination.

The selection of indicators will already incorporate some of the presuppositions of the ranking compilers as to what counts as a 'world class university'. That concept might include a university that has a worldwide reputation, produces pivotal research, has an international and cosmopolitan staff and student body or produces well qualified graduates that are equipped to play a leading part in the world of commerce and industry. Each of those various ideas of what constitutes a world class university will prompt a different selection of indicators.

Nor can it be assumed that because a measure is objective, it has the same meaning in all cultural contexts. The proportion of members of staff, teachers and researchers, who come to the institution from overseas may mean different things in different contexts. Quite apart from anything else, 'overseas' involves rather different concepts of scale in the cases of Singapore and the United States. In a university which sees its mission as being to serve a region that extends beyond national boundaries, as the National University of Singapore or the University of Cape Town may do, internationalism will play a different part in institutional culture than in a university that sees its mission as being to serve a region that falls entirely within a single country.

From the possible range of indicators that could be chosen, a fairly narrow range of indicators actually are chosen. They are chosen generally on the basis that they are relatively cheap to collect. However, what those indicators actually measure, the question of the validity of the measures, is open to serious question.

METHODS OF ANALYSIS

Unfortunately, the selection of relevant measures for calculating the rankings is only the start of the arbitrary decisions the compilers of league tables make. Much more important, in many ways, are the methods that are used to calculate rankings. In order to keep those 2 elements separate, at least for the time being, in this section I will use the data from the Shanghai Jiao Tong University (SJTU) rankings for 2008, so that it is clear that the nature and quality of the data is not the focus of attention.

Six measures were used in the calculation of the table in 2008. There are minor changes in the way that tables are calculated from year to year which do not need to concern us here either; the method has remained fairly constant over time. Those 6 measures are shown in Table 1.

Before examining how the table compilers actually put their rankings together, let us consider the data for a moment, and think about the educational processes involved. There is a mathematical approach that is ideally suited to examining the data in cases such as this, where different units have a different focus or mission, and where there is no clear-cut standard by which efficiency can be judged. That method is Data Envelopment Analysis (DEA) (Emrouznejad 1995–2011). I will use a very simple analysis to illustrate how it works, and then I will use a purpose-designed computer programme to carry out the analysis more fully.

From the data provided by SJTU, I have calculated 3 variables in order to simplify the illustration. I have taken the 'Quality of Education' assessment criterion as it is. For each of the criteria, 'Quality of Faculty' and 'Research Output' I have taken the 2 components and found the average. And finally I have, for the moment, ignored 'Per Capita Performance'.

If we think for a moment about the process involved in a university, the university authorities, or the government agency that funds the university, provide high quality, trained members of staff, and in return expect the university to produce well qualified alumni and alumnae, and world class research publications. A world class university, therefore, might be expected to be world class at converting that input into those outputs. In order to give a sense of that, I have divided each of the indicators 'Quality of Education' and 'Research Output' by the measure for 'Quality of Faculty', as an indication of how productive the employees are in terms of teaching and research output, respectively. Plotting a graph of those 2 resulting variables produces Fig. 1. Teaching output per unit input is plotted on the vertical axis, while research output per unit input is plotted along the horizontal axis. Only the top 200 universities from the SJTU ranking have been plotted.

Three straight lines have been added to the graph to indicate the limit to the data. Institutions that become better at producing outputs will move upwards and to the right on the graph, and the limit that has currently been reached is defined by the 3 institutions that have gone furthest in those directions. Those institutions which are 'pushing the envelope' define the data envelope that gives its name to this method.

Incidentally, it will be noticed that there are a lot of institutions that lie along the horizontal axis on the graph. This is because they all scored zero on the 'Quality of Education' output. That is not very surprising in the sense that quality of education was

Table 1. Indicators used in the Shanghai Jiao Tong University Rankings for 2008. Source: www.arwu.org/ARWUMethodology2008.jsp, where details of the definitions used are available

Criterion	Indicator	Code	Weight
Quality of Education	Alumni of an institution winning Nobel Prizes and Fields Medals	Alumni	10%
Quality of Faculty	Staff of an institution winning Nobel Prizes and Fields Medals	Award	20%
	Highly cited researchers in 21 broad subject categories	HiCi	20%
Research Output	Articles published in Nature and Science	N&S	20%
	Articles indexed in Science Citation Index Expanded, and Social Science Citation Index	PUB	20%
Per Capita Performance	Per capita academic performance of an institution	PCP	10%
Total			100%

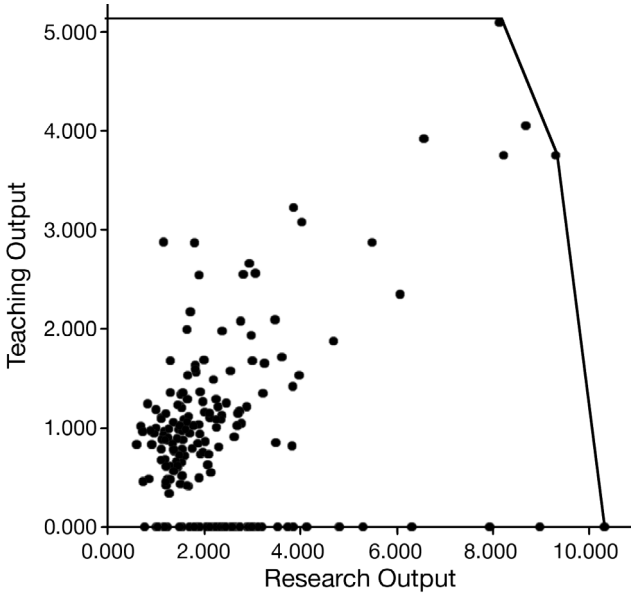

Fig. 1. Research and teaching output of world class universities

measured entirely in terms of alumni who had been awarded the Nobel Prize or Fields Medal. We would perhaps think an institution strange if it evaluated its success as a teaching institution entirely in those terms. However, the selection of measures is not at issue at this point.

The 3 institutions that are on the data envelope, starting on the horizontal axis and moving anti-clockwise, are Seoul National University, the National University of Taiwan and the University of Vienna. The standard procedure in DEA is then to give each institution a score based on how close they are to the part of the data envelope that they are closest to. Institutions that are on the envelope are given a score of 100, while other institutions that are closest to the envelope receive a high score. In this case, the University of Barcelona (just to the left of Seoul National University), and Delft University of Technology receive an honourable mention. But it is the Universidad Nacional Autonoma de Mexico that comes closest to the data envelope and will therefore have the highest non-100 score. It is, perhaps, worth commenting that none of these institutions is in a country where the official language is English.

Harvard, with a score of 1 on each of the axes, is somewhere in the middle of the clump of run-of-the-mill institutions, as are most of the top 20 institutions in the SJTU ranking.

This illustration used only 1 input and 2 outputs, not because that is a restriction imposed by DEA, but because that is a limit imposed by a visual image drawn on a 2-dimensional sheet of paper. DEA is capable of dealing with as many inputs and outputs as one wishes. So now, going back to the 5 indicators used by SJTU (again ignoring the size of the institution) we can use a commercially available DEA programme to analyse the data. I have used Frontier Analyst Professional Version 3.1.5, produced by Banxia Software.

In the next section, I will discuss the results of DEA analysis.

DEA ANALYSIS

DEA was developed for comparing 'decision making units' which have a different mix of activities, with the result that direct comparisons are complicated. For example, a number of motor factories might be compared, but each factory might produce a different combination of cars, small vans and SUVs. Or we might wish to compare the performance of supermarkets, but a supermarket in a suburban area might have a different mix of clientele than one in a rural area, so they would have different patterns of performance. As in the illustration above, DEA acknowledges that there may be many ways of being excellent, as indicated by different points on the data envelope.

Each institution that is not on the data envelope is then presumed to be aspiring to reach the nearest point on the data envelope. In effect, that means that the institutions that it is compared with will be those that are most similar in the mix of activities that they undertake. That gives DEA 2 major advantages over other methods of analysis, namely that it takes into account the fact that some performance indicators measure inputs while others measure outputs, and it ensures that like institutions are compared with like.

Having said that, deciding whether a performance indicator measures an input or an output may not be uncontentious. For example, I have suggested above that the number of faculty who have been awarded Nobel Prizes is an input, on the grounds that those scholars may be expected to stimulate future research work by their students and colleagues, and are therefore part of the environment of scholarship that the university provides. In contrast with this view, a university administrator might argue that they have been able to attract a large number of Nobel laureates to work in their university, because of the excellent reputation of the institution. They might, therefore, see this as an indicator of the repu-

tation of the institution, and thus as representing an output.

I have elsewhere (Turner 2005) noted that similarly conflicting arguments can be made for almost any indicator depending upon one's perspective. For example, among applicants to a university, high entry qualifications may be seen as an input, which the students bring to the educational process. On the other hand, high entry requirements can also be seen as an indicator of reputation and esteem, and therefore be seen as an output that applicants seek.

Testing the robustness of results in DEA is difficult because there are no standard statistical tests, and because the method is fairly sensitive to the inclusion or exclusion of particular institutions, especially if those institutions are on the data envelope. The only effective way of testing the robustness of DEA results is to run many analyses and compare the results to see which alterations have a big impact on them. For that reason, DEA is more useful when it is used, as in this present case, to explore the range of possible modes of analysis, rather than when hoping to produce a definitive ranking.

However, in order to extend the illustration further, some decisions have to be made about which indicators to treat as inputs and which as outputs. I will, in this case, continue from the perspective of the previous analysis, and treat 'Quality of Faculty' as an input, and the other indicators as outputs. However, in this case, I shall disaggregate the 2 indicators that made up 'Quality of Faculty' and 'Research Output', giving 2 input variables and 3 output variables.

Taking the top 200 institutions from the SJTU ranking, and using those variables and inputs and outputs, the results of the DEA analysis are as shown in Table S1 in the Supplement at www.int-res.com/articles/suppl/e013pp1_supp.pdf. One of the risks with increasing the number of inputs and outputs is that there are more ways for an institution to excel, so that if one includes enough input and output indicators, eventually all that one demonstrates is that each institution is special in its own way. Even by increasing the number of inputs only modestly, to 2 inputs and 3 outputs, no fewer than 26 institutions, of the top 200 from the SJTU ranking, are on the data envelope.

DEA can also offer much more in terms of detailed analysis than simply producing a simple ranking. Each institution is compared with a notional 'ideal' institution which is made up of aspects of the most effective institutions combined to give the same proportions of inputs and outputs as the institution under consideration. To take a concrete example, we might ask why Harvard faired so badly in the DEA analysis. The first step to answering this is to find out which institutions it was being compared with. Fig. 2 shows the output from Frontier Analyst, showing how the comparator for Harvard was made up.

As can be seen from Fig. 2, Harvard is compared with a composite that incorporates the characteristics of Johns Hopkins University, Moscow State University and Delft University of Technology, but by far the biggest component comes from Johns Hopkins University. If, then, we compare the performance of Harvard and Johns Hopkins, as shown in Fig. 3, we can see that Harvard clearly has a higher score on all the indicators than Johns Hopkins. However, the score on HiCi and Award (and especially on Award) is disproportionately higher, suggesting that Harvard has a disproportionate focus on very high status research, compared with the overall level of research activity in the institution.

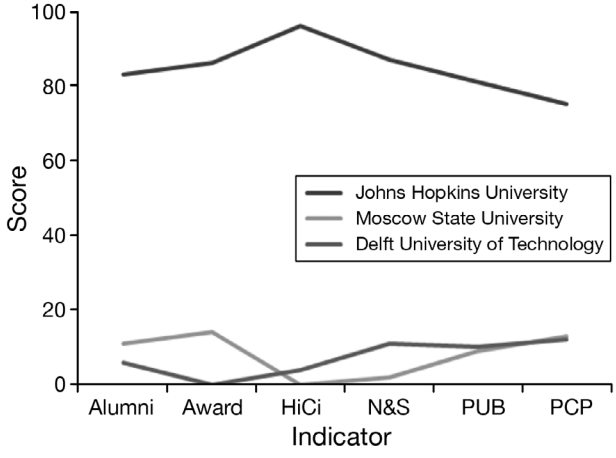

Fig. 2. Institutions used as reference for Harvard University. See Table 1 for codes

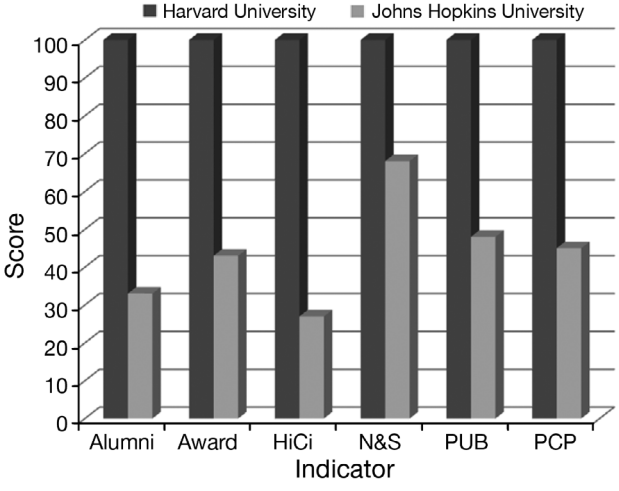

Fig. 3. Comparison of the scores of Harvard University and Johns Hopkins University. See Table 1 for codes

Of course, there is room for debate as to whether this is a correct perception, and that there should be a particular balance between accomplished scholars and up-and-coming scholars in a world class institution. Should a world class university simply be, in effect, a research institution where scholars go once their achievements have been recognised, or should a world class university have a role in developing new researchers, and therefore include a proportion of scholars who have yet to achieve recognition? And if the latter, what would be an appropriate mix of the different kinds of activities? These are important questions, but are unlikely to be addressed if administrators and policy makers chase after places in the rankings without questioning what it means for the activities that universities undertake.

In terms of international comparisons and the cultural bias of league tables, the position of the University of Vienna is interesting. It also performs atypically in the *Times Higher* – QS (2008) rankings. For some reason, it also has a very poor score on the key input in that ranking, the Student/Staff scores. And because it scores very poorly on an input, the outputs produced look disproportionately more impressive. Of course, in this context, the data alone cannot tell us what that reason might be for the University of Vienna to be an outlier. It is possible that the data is being collected on a different basis in different countries, and that, as a result, the data is simply unreliable. But it is also possible that there are cultural and historic reasons why the data should represent systematically different things in different countries, and indicate that the whole project of producing world rankings is either much more difficult than it seems, or completely pointless.

In relation to the University of Vienna, there are 2 important differences between the traditions of continental Europe and those of the Anglo-Saxon world which would make this very singular result for an Austrian university explicable. In the first place, it has been usual practice in many continental European countries, and especially German-speaking countries, for teachers in the elite secondary schools (gymnasia) to work part time in a university while completing their studies for a higher degree. This made for much more fluidity of employment between those high schools and posts as junior members of faculty in the universities than is common in an Anglo-Saxon context (Teichler & Bracht 2006). That might result in some disparities in how the members of staff are counted in different cultural contexts, and an unrealistically low figure for an institution in Europe.

The other major difference arises from different admission practices and patterns of attendance. Whereas, in Anglo-Saxon institutions, entry requirements are high, and the expectation is that all or most students who enrol will complete the programme of study, in continental Europe, the expectation was that every student who satisfactorily completed upper secondary school had the right to attend the university/course of their choice. This made for some very large, indeed unmanageably large, entry level classes, and high drop-out or cool-out rates. The idea that entry might be restricted to only those students who achieved the best marks, and that class sizes might legitimately be limited in order to maintain quality, came relatively late in many national systems of higher education in Europe (European Commission 2006). In any case, since fees were low, rarely more than a peppercorn, it made sense to drop in and drop out of courses of study, and to spread study over a longer time span. In that way, students could compensate for poor quality experiences of lectures and seminars that were overcrowded by private study, and drop back in the following year when they were better prepared to cope with difficult study conditions. I have even heard it suggested that the fees were so low, and the discounts for students in restaurants, shops and cinemas so good, that it made economic sense to enrol as a student, even if one had no intention of studying anything. All of these practices would lead to an unrealistically high figure for the number of students who were seriously following a course of study.

It is interesting to note that many of the 'quality indicators' that have been adopted in national league tables in the UK, US and Canada, and that are influencing notions of quality worldwide, run directly counter to this long-standing tradition of higher education; continuation rates, final qualification outcomes and proportion of students completing within a specified time have all been used as quality measures in the UK, but would have produced incomparably different figures for, say, a traditional English university and a traditional French one.

Different cultural contexts will apply to different historical traditions in higher education, and there may be a variety of explanations as to why institutions in English-speaking countries do very well in the SJTU rankings and *Times Higher* – QS rankings, but show up rather poorly in DEA analyses that take some measure of inputs into account. However, this would certainly seem to be an area that would benefit from further investigation.

ACTUAL RANKING SYSTEMS

Having used this brief tour of DEA to highlight the range of choices that have to be made when developing a ranking system, to select indicators, to select methods of analysis, to choose whether to treat indicators as inputs or outputs and so on, it is now time to consider the ranking systems that are actually given credence when thinking about world class universities. To continue Birnbaum's (2007) metaphor, now that I have described the Library of Babel, and the infinite variety of possible ranking systems that it contains, it is time to consider those examples that have been picked out by scholars in the library as making sense. There are really only 2 major, international ranking systems that need to be considered, SJTU and the *Times Higher* – QS.

On the face of it, the SJTU rankings are constructed on a quite different basis from those of the *Times Higher* – QS. While the former depend on measures of high status research for at least half of their weighted score, the latter use measures that indicate the status of the institutions in the eyes of professionals closely related to academia. Again, it cannot be stressed too often what a very narrow range of high status research the SJTU rankings depend on; having even a single member of faculty who was a Nobel laureate would have a significant impact on an institution's position, although whether the person concerned has to be a member of faculty at the time of the award, or can be bought in after the event, is unclear from the details of the methodology published on the website. In contrast with that, the *Times Higher* – QS rankings attempt to take the broadest possible sampling of opinion across all continents. Although that effort is doubtless constrained by resources, the 2007 rankings included the judgements of over 6000 respondents.

Notwithstanding those differences, and other technical differences between the rankings, which make it difficult to compare them directly, the 2 rankings look startlingly similar. (SJTU rate many medical centres separately from their parent institutions, for example, whereas it is not clear from the *Times Higher* – QS table how they treat those centres, although it seems likely that they are included with the ranking of the parent institution overall. Similarly, some groupings of institutions, such as the University of California system, are not divided in the same way in the 2 tables.) Both tables have Harvard, Yale, Stanford, Cambridge and Oxford near the top, while the ranking maintains the pecking order that one would expect to find in the corresponding national league table. Those UK institutions that appear are those that come near the top of the league tables that are published by the *Times*, *Guardian* and *Independent*; Bristol, Edinburgh, Durham, Manchester and Newcastle all feature near the top. Similarly, Japanese institutions appear in a characteristic and hierarchical pattern that is familiar to scholars who know the system, with Tokyo and Kyoto Universities at the top, followed by Nagoya, Hokkaido, Hiroshima and Waseda.

What systematic differences there are between the 2 tables point to the difficulties of constructing such league tables on a cross-cultural basis. Having an emphasis on medicine or the physical sciences confers a huge advantage on institutions in the SJTU tables, which arises from the choice of Nobel prizes as an indicator, Nobel prizes themselves having an unequal distribution across the disciplines. The *Times Higher* – QS table treats francophone institutions, and others that do not use English as the primary medium of instruction, rather more favourably than the SJTU table. But again, that can be readily understood in terms of the use of reputation, which can relatively easily cross language barriers, compared with citation counts in indexes that focus on English language publications. These differences arise from distinct cultures, whether language based, nation based, or discipline based, that can also be seen reflected in the national league tables, although the extent of variation is clearly much greater than in the case of national comparisons.

Notwithstanding all of those differences in detail between the 2 tables, 334 institutions that appear in the Jiao Tong top 500 institutions also appear in the *Times Higher* – QS top 400. Picking out only those institutions common to both tables, and creating a ranking of them depending on where they are in each of the league tables, produces 2 new rankings. The Spearman's rank order coefficient of those 2 scales is 0.68, showing a very strong correlation. Great caution needs to be exercised in interpreting this figure. Since it clearly ignores the institutions that appeared in one table but not in the other, it ignores those cases where the divergence between the 2 tables was the greatest, and will therefore be an overestimate of the correlation.

Many of the institutions that appear in only one table were medical centres, which appeared in the SJTU but not in the *Times Higher* – QS table, as discussed above. In addition, fewer than 30 institutions that SJTU ranked between 400 and 500 appear among those 334 institutions. For the most part, therefore, we are looking at a correlation between

over 300 of the institutions that appeared in the top 400 of the 2 tables. The statistical problems may therefore not be quite as severe as they appear initially.

On the other hand, the task faced by the compilers at SJTU and the *Times Higher* – QS was not to rank a predetermined group of 334 institutions. Their task was to select the 'best' institutions from a pool of literally thousands of candidates, a task with many more degrees of freedom than ranking a preselected group. In that sense, the Spearman's rank order coefficient may seriously underestimate the coincidence of those 2 rankings.

For those reasons, I would not wish to attach any particular significance to the actual number produced by a comparison of the 2 rankings, but offer it only as qualitative evidence, such as it is, that there are strong similarities between the 2 sets of outcomes, in spite of the quite marked differences in underlying approach to the indicators chosen. Those who are less inclined to trust statistical methods can cast their eye over the original tables, and will probably form a similar impression; the same institutions turn up in roughly the same position on both.

The question that needs to be addressed, therefore, is why, from all the possible ways of constructing league tables that could have been chosen, have these 2 actually been chosen, and why do they give such similar results? Again, DEA can be helpful in examining the question. If there is a set of assumptions about inputs and outputs, which, when combined with DEA, produces a rank order similar to SJTU or the *Times Higher* – QS tables, then that will tell us a lot about the implicit assumptions behind the rankings that have been selected.

As it happens, there is a set of assumptions about inputs and outputs that when used with DEA produces a very similar outcome to the original SJTU ranking. If all of the performance indicators are treated as outputs, and if, in addition, a dummy input variable is created and given an arbitrary value of 100, then the ranking is very similar to that of the rankings of the SJTU and *Times Higher* – QS. Table 2 compares the top 20 institutions obtained by SJTU and DEA using that assumption.

This result is really not very surprising. The standard technique in most league tables is to select a range of indicators, assign an arbitrary weighting to each, and then simply add all of the weighted indicators. This, in effect, treats all the indicators as outputs; any increase in any indicator moves the institution up the ranking, although by how much depends upon the weighting. It also effectively ignores any inputs as irrelevant. Rich institutions can pour any amount of resources into the processes, and not be penalised for it. Whereas in DEA each institution is compared with those effective institutions that it most closely resembles, in the currently used rankings each institution is compared, and not very favourably, with Harvard.

Table 2. Comparison of Shanghai Jiao Tong University (SJTU) ranking with Data Envelopment Analysis (DEA) ranking with restricted assumptions

SJTU top 20	SJTU score	DEA top 20	DEA score
Harvard University	100	Harvard University	100
Stanford University	73.7	University of Cambridge	91.5
University of California - Berkeley	71.4	Stanford University	86.6
University of Cambridge	70.4	Tokyo University	83.3
Massachusetts Institute of Technology (MIT)	69.6	University of Chicago	81.9
California Institute of Technology	65.4	Massachusetts Institute of Technology (MIT)	80.6
Columbia University	62.5	Princeton University	80.4
Princeton University	58.9	University of Toronto	79.8
University of Chicago	57.1	University of Michigan - Ann Arbor	78.1
University of Oxford	56.8	University of California - Berkeley	77.1
Yale University	54.9	University of California - Los Angeles	75.7
Cornell University	54.1	University of Washington - Seattle	74.1
University of California - Los Angeles	52.4	Columbia University	72.4
University of California - San Diego	50.3	Kyoto University	70.6
University of Pennsylvania	49.0	California Institute of Technology	69.1
University of Washington - Seattle	48.3	University of Pennsylvania	69
University of Wisconsin - Madison	47.4	University of Sao Paulo	69
University of California - San Francisco	46.6	Johns Hopkins University	68.5
Tokyo University	46.4	University of Wisconsin - Madison	68.1
Johns Hopkins University	45.5	University of Minnesota - Twin Cities	67.5

We should not be surprised that if we view world class universities as being those that come out on top in the traditional rankings, we find they are equivalent to rich institutions with large endowments. If you want to come out on top in the rankings that have been widely used to define world class, the solution is simple; throw money at the problem. Nor should we be surprised that the World Bank comes to the conclusion that a major element in the generation of world class universities is an abundance of resources (Salmi 2009, p. 8). If we looked at those institutions that make the most effective use of limited resources, we might have a very different view of what makes a world class university.

There remains a question about why apparently diverse indicators, in the case of SJTU and the *Times Higher* – QS rankings, gave similar results. Proponents of the ranking systems would doubtless argue that this indicated that they had found some underlying construct of 'quality' that they were measuring. Our exploration of the variety of outcomes available through DEA should discourage us from taking that simplistic view. Rather, it seems likely that most of the indicators are a proxy in one way or another for the reputation of the institution. The *Times Higher* – QS survey definitely tried to measure reputation directly, and it seems reasonable to think that academics who publish in respected journals and win international prizes might well be found in prestigious universities.

However, the exact mechanism by which that operates is not illuminated by the statistics alone. It may well be, as the compilers of the SJTU ranking would wish us to believe, that articles published in very prestigious journals contribute to the reputation of the university. Or it might be that academics who can publish in prestigious journals are attracted to universities with strong reputations. But it is just possible that academics who would, in other institutions, be regarded as mediocre, are able to publish in prestigious journals because they can rely on the reputation of the university where they work. Reputation, by its very nature, is a slippery concept that tends to colour perceptions of other aspects of institutional life. A good deal more work is needed to find out exactly how these elements work together to create a world class institution.

CONCLUSIONS

In this paper, I have introduced the idea of using DEA as a method for managing comparisons between institutions at an international level. DEA is a widely used analytical approach, and is accepted in engineering, logistics, economics and many other fields as a valuable way of comparing the performance of units where there is no clear-cut and unambiguous way of measuring effectiveness. In the field of education, and benchmarking educational institutions, however, it is virtually unknown; perhaps it would be better to say 'unused', rather than 'unknown'. A report for the Higher Education Funding Council for England considered the use of DEA and concluded that, 'DEA provides for each institution an efficient reference set of peers, which would seem to be useful for benchmarking. However, some of the reference institutions were, to anyone with any knowledge of the sector, completely inappropriate' (Performance Indicators Steering Group [PISG] 1999, p. 47). Which is to say, we know that Johns Hopkins University is not an appropriate reference institution for Harvard, because Harvard always comes out at the top of any ranking.

There is the obvious danger here that any future ranking system will need to demonstrate that it conforms with what 'anyone with any knowledge of the sector' would think. Further development of ranking systems would therefore appear to be pointless, since if they tell us anything different from what we know today, and therefore different from what we are told by the present ranking systems, we will be obliged to ignore it. Even today, systems that rank educational institutions may be nothing more than a way of quantifying what we want to believe in the first place, and thereby add a veneer of scientific respectability and quantification to our well-worn prejudices.

I have not, however, in advancing DEA, been primarily interested in supplanting one ranking system with another. The main interest in DEA, as far as I am concerned, is that it raises questions about what counts as quality, and could be used to frame debate about how we recognise and reward quality in higher education. By demonstrating that current ranking systems focus on outputs at the expense of all else, DEA helps to frame questions about whether that is a desirable emphasis. Harvard, Princeton, Oxford and Cambridge are wonderful institutions in their own way, and produce research of unimpeachable quality. But at what cost? In a period where higher education worldwide is likely to be expected to produce more for less, might it not be better to look for role models in Leeds, or Singapore, or Vienna, rather than always following the lead of the usual suspects?

These are questions that cannot be answered in a single paper, but they should at least be asked. We look to universities, and university research, to produce the big ideas that will shape our futures. Government policy makers, and even academics themselves, are not always very good at spotting where the next big idea is coming from, or even recognising it when it has arrived. Until we have some indication that the assumptions of policy makers are based on evidence, we should be sceptical of proposals to invest all of our research prospects in a single basket of eggs made up of a small group of world class universities.

LITERATURE CITED

Birnbaum R (2007) No world-class university left behind. Int High Educ 47:7–9

Borges JL (1970) Labyrinths: selected short stories and other writings. Penguin Books, Harmondsworth

Emrouznejad A (1995–2011) Data Envelopment Analysis homepage. www.DEAzone.com (accessed 9 February 2013)

European Commission (2006) The extent and impact of higher education governance reform across Europe: Final Report to the Directorate-General for Education and Culture of the European Commission. Center for Higher Education Policy Studies (CHEPS), University of Twente, Enschede. http://ec.europa.eu/education/pdf/doc236_en.pdf (accessed 9 February 2013)

Nobel Foundation (2009) All Nobel Laureates. Nobel Foundation, Stockholm. http://nobelprize.org/nobel_prizes/lists/all/ (accessed 1 September 2009)

Performance Indicators Steering Group (PISG) (1999) Performance indicators in higher education: First Report of the Performance Indicators Steering Group (PISG). Higher Education Funding Council for England, Bristol

Salmi J (2009) The challenge of establishing world-class universities. The World Bank, Washington, DC. http://siteresources.worldbank.org/EDUCATION/Resources/278200-1099079877269/547664-1099079956815/547670-1237305262556/WCU.pdf (accessed 13 April 2010)

Shanghai Jiao Tong University (2008) 2008 Academic ranking of world universities. Shanghai Jiao Tong University, Shanghai. www.arwu.org/ (accessed 1 September 2009)

Teichler U, Bracht O (2006) The academic profession in Germany. In: Hiroshima University (2006) Reports of changing academic profession project workshop on quality, relevance, and governance in the changing academia: international perspectives. Research Institute for Higher Education, Hiroshima, p 129–150

Times Higher – QS (2008) The top 200 world universities. www.topuniversity.com/university-rankings/world-university-rankings/2008 (accessed 25 April 2013)

Turner DA (2005) Benchmarking in universities: League tables revisited. Oxf Rev Educ 31:353–371

Editorial responsibility: Konstantinos Stergiou, Thessaloniki, Greece

Submitted: December 18, 2012; Accepted: February 14, 2013
Proofs received from author(s): April 17, 2013