Chapter 5
Ranking Indicators and Weights

Bernard Longden

5.1 University Ranking: Reliability, Consistency and Validity

> The standard market research techniques followed in data collection have raised the question: Can we assume that what works for pet food, perfume, and pesticide will also work for education? (Stella and Woodhouse 2006: 10)

5.1.1 Introduction: Positionality and Ideology

University rankings are ubiquitous and here to stay, but they are a feature of the contemporary higher education agenda. Harvey (2008: 187) reminds us that the ascendency of league tables in the higher education agenda has much to do with the 'increasing marketisation of higher education, greater mobility of students, and ultimately the recruitment of foreign students.'

The position held by stakeholders, on the worth and value of university rankings is diverse. Given the potential for a polemical position on this worth and value, it is deemed important that the ideological position for this chapter is made clear and unambiguous from the start.

A strong position is taken by Brown (2006: 38), who remains fundamentally opposed to any support of commercially produced university ranking or league tables. The basis of his criticism rests with the claim by publishers that they address matters related to quality of university teaching and research. In profoundly

B. Longden (✉)
Professor of Higher Education Policy, Liverpool Hope University, Liverpool, L16 9JD, UK
e-mail: Bernard.longden@gmail.com

J.C. Shin et al. (eds.), *University Rankings*, The Changing Academy – The Changing Academic Profession in International Comparative Perspective 3,
DOI 10.1007/978-94-007-1116-7_5, © Springer Science+Business Media B.V. 2011

challenging their proposition, he has identified four arguments to support his opposition; they are:

- Rankings are based on data and assumptions about data that are scientifically questionable
- University rankings will influence universities to produce the 'wrong' kind of higher education
- League tables reinforce the tendency to see higher education as a product to be consumed rather than an opportunity to be experienced and viewed as being 'just another commodity'
- Risk of allowing commercial considerations inexorably leads the university to a position where the market determines quality. More generally, the creation of the impression that some institutions are better than others when in a diverse, mass system there can be no one 'best university' or single view of quality. League tables indeed strengthen the market position of institutions that are already prestigious and well funded, at the expense of those that may be seeking to build reputation by attending to the needs of students and employers

Aligning to this position is to conclude that rankings misrepresent the work of universities and colleges in the interest of selling newspapers.

While this strong line of argument is attractive, there is a risk of 'tilting at windmills' which may not be so productive. In a report to the Standing Conference of Principals, my colleague Mantz Yorke concluded our report with an acknowledgement that, while they have serious limitations, it is better to work to improve them and as Lennon and McCartney say in the words of Hey Jude, '… take a bad song and make it better.'(Yorke and Longden 2005: 35).

This scepticism, which has been outlined above, remains powerful and has influenced the internal logic to this chapter and therefore this chapter reflects an ideology sympathetic to this scepticism.

5.1.2 It Starts with Events

Much of our daily routine relies on data. For example, your morning ritual may involve standing on the bathroom scales, looking at the readout, knowing that while it may not give the weight desired, it displays, accurately, numbers which are consistent and reliable. The readout from a bathroom scale is neither spurious nor idiosyncratic – it has a meaning.

The bathroom scales is a *black box* – a metaphor.

Concerns about the internal working are of limited interest as long as the readout remains accurate – accurate to a degree that is suitable for personal needs. Most of us are disinterested in finding out about how it all works, or what assumptions and adaptations have been applied.

In the physical world, we secure and gain confidence about the various pieces of physical equipment through experience. In the social world, we are dependent on trust.

An assessment about the reliability and the esteem of the person providing answers to our questions is critical to gaining that confidence.

Confidence that the metaphor of the bathroom scales is underpinned by theory is sufficient for us to entrust it.

Demands on our busy force us to place a degree of trust in the many *black boxes* we come across each day. Opening the *black box* each time, we need to be reassured that what we are dealing with s valid, reliable, and consistent, which is not a realistic possibility.

Life would become impossible without some degree of confidence and trust in the validity of the information that the physical or social world generates.

The concept of trust becomes an integral part of the human condition; without it, our every experience would become a series of hypothesis open to rigorous testing before decision could be made about the experience.

5.1.3 Locating a 'Black Box'

The thesis, underpinning this chapter, is that compilers of university rankings rely on us not opening the *black box*; not to question the interrelationships between the various parts within the *black box*; and not to question the construction of the various elements within the *black box*.

Constructing the university ranking relies on a conspiracy between the compilers and the wider community that the *black box* remains unchallenged. The mathematical and statistical complexity embedded within the *black box* would demand too much valuable time and effort to unpick, leaving the reader of the rankings depend on a trust in those who construct the rankings; after all, the reader is probably only interested in the final column – the aggregate sum of the other columns shown in the table.

The final ranking index provides some form of objective measure – a measure that, in some real sense, maps an underlying set of characteristics that, when aggregated in a particular way, provides a ranking.

Few are prepared to dig beneath the top line index to find out how the final measure was achieved – how the *black box* actually works. Compilers remain confident that it is unlikely that serious criticism will come from an analysis of the content of the *black box* because of the 'challenging mathematical and statistical complexity' that inevitably would ensue.

By negating a willingness to open the *black box* and peer inside, we have handed over to the compilers a consent that what they have done is acceptable – that they have provided a reliable, consistent, and valid processes that measure or reflect the 'worth' of the universities ranked.

So, to open the *black box* requires effort and a sense of a critical commitment to understand how the mechanism works. For most of the time we simply want a quick answer to the question 'which is the best university?' as if such a deceptively simple question can mask the complexity beneath.

Table 5.1 Problems intrinsic in designing university rankings

1. Become an end in themselves protected from critical scrutiny
2. National, institutional and program diversity
3. National and linguistic diversity
4. Partial coverage of purposes and stakeholders
5. Problems of aggregation and weightings
6. Reputational rankings
7. Produced context-free judgements
8. Undermine universal improvements
9. Reduce scope for innovations in strategy, curriculum, pedagogy and research

Other chapters provide a more detailed discussion on the use and abuse of university rankings; however, it is appropriate at this point to be reminded of some of the main problems associated with the process of ranking universities.

Table 5.1 provides a list of some of the intrinsic problems that university rankings generate. In this chapter, the focus will be on the methodological issues that arise during the creating of a university ranking index, although some of the other problems identified in Table 5.1 may be addressed in passing.

There remains a challenge that when the *black box* is opened is it possible to understand the processes deployed by the compilers? Do compilers ensure satisfactory levels of transparency in describing their methodologies? Is there a commonality to the methodologies that different compilers employ in constructing the ranking index? How visible is the internal mechanism of the *black box*?

I propose to focus on these questions to gain an insight into the problems associated with aggregation and weightings of performance indicators in Table 5.1. Along the way, I will briefly address other issues listed in the table but the main objective will be the focus on the key confidence measures or validity, reliability and consistency as these remain the cornerstones of the trust that is given to those compiling the university rankings.

5.2 Critical Steps and Economical Truths

A university ranking index provides an end point for the user by a process of consolidating a large data set, a single index that in some ways represents the ‘university’. This simple statement exposes the facile nature of the process. How can the activity of a university be reduced to a numerical value? Anyone who works within a university setting knows only too well that within the one institution, there are pockets of high quality and pockets that are of concern to the institution. Providing a single measure betrays the complexity of the institution. Unlike a car manufacturer where there is a product line to measure, universities have different aims from each other and therefore comparisons that fail to take note of the differing ‘missions’ fail to make sensible comparisons.

Clarke (2002, 2004) describes two common criticisms relating specifically to *US News* and *World Report* and the methodology used to rank colleges and universities. First, the compilers constantly change the formula they use to create the rankings and thus make the interpretation of yearly shifts in a university/college ranking, in terms of academic quality, impossible. Second, the score used to assign schools to ranks is overly precise, creating a vertical column where a group might more properly exist.

In this section I plan to tease out aspects of the mechanisms within the *black box* to ensure that there is an understanding of the techniques deployed. These techniques need to be understood so that assumptions implicit within the process can be appreciated and create a more transparent methodology capable of evaluation by the user. There are a number of specific elements within the methodological *black box* that will be considered:

- Selection of indicators to produce the final ranking index, issues surrounding the way in which indicators are added together
- The relative weighting that is applied to the various indicators deployed to create the final ranking index
- Management of missing data
- Statistical differences between the ranking indices that emerge

Seven steps can be identified clearly in the process of creating a university ranking; consideration of each of those stages follows.

5.3 Steps Towards Creating a University Ranking Index

The problem with ranking concerns the practice not the principle. (Altbach 2006).

How is it possible to accurately reduce a university performance to a single index? A university is a cauldron of beliefs, values, and actions and the proposition that it could be possible to distil this all down to a single index remains, for me, a challenge and a fear that there may be more reliance on alchemy than on logic.

However, it would be inaccurate to suggest that compilers involved in preparing and publishing university rankings seek to keep the box tightly closed. Quite the reverse, most seek to provide the reader with a very detailed account of how they compile the ranking indices (Morse and Flanigan 2006; O'Leary et al. 2006; MacLeod and Hiely-Rayner 2007, 2008, 2009a). Recently, the *Times Higher Education* has been at pains to make adjustments to the methodology they use in calculating the 'world rankings' (Baty 2009).

In considering the stages necessary in producing university rankings, it is possible to identify key processes that all compilers appear to adopt. For the purpose of this chapter, I propose to explore each of these stages in some detail. The starting point of the process is the measurement of an event that relate directly to the university activity. The measurement or performance indicator (PI) – is ubiquitous, often it is invisible, which helps define the institution. When aggregated with other

Table 5.2 Critical steps that pose potential difficulties when creating a university ranking

1. Clarifying reason for ranking
2. Selecting suitable metrics – performance indicators (PI)
3. Collecting data – metrics
4. Adaptation of PIs into a scale
5. Standardising measures prior to aggregating
6. Weighting PIs prior to aggregating
7. Creating a single index reflecting a university

measures, it can provide a numerical shorthand for key characteristics about the university.

The steps deployed need to be identified and confidence need to be secured so that assumptions, adaptations and definitions are fit for purpose. Using the critical steps as a guide, it is possible to show points where potential difficulties can occur and are often overlooked or ignored when providing a narrative on how rankings are created.

Teasing out these critical steps provides a means by which those elements of the process that are vulnerable to mystification and obscuration can be considered in detail (Table 5.2).

5.3.1 Clarifying Reason for Creating University Ranking

Altruism is unlikely to be the justification advanced by a publishing company engaged in producing a university ranking table. The reason why publishers involved in this genre of publishing retain their involvement is simply down to the money they generate from the final product – advertising revenue, purchases of the final ranking book and other forms of endorsement. The really great thing for the publisher is that once the template for the production of the tables has been established each year, a new target population is ready to buy their product.

Two main types of university rankings are evident in the commercial world of rankings. The audience for the two types of university rankings is distinct and different but the methodology adopted by compilers to create the rankings is very similar.

5.3.1.1 Type 1: Undergraduate Experience: Teaching

US News and *World Report* (*USA*), *Maclean's* (Canada), *The Guardian* (*UK*) and *The Times* (UK) all have as their target audience the potential undergraduate student market. The common feature of all these and other similar publications is the production of a ranking that, it claims, reflects the quality of teaching and learning within higher education institutions. The measures that are used to reflect this quality index are those that relate directly or indirectly to the undergraduate experience.

The complexity of the data that is required to produce comparisons is such that when coupled with the diversity of provision across countries, the university rankings tend to be specific to a country. Attempts at providing the global rankings for teaching and learning have remained elusive so far. The main reason for this must relate the need to secure a common set of definitions for the measures employed in the creation of the rankings. Given the diversity of provision this remains an obstacle.

The nature of measures typically associated with domestic university rankings would be student staff ratios, spend per student FTE, student satisfaction measures. Even within one country, the diversity of measures or PIs used in the calculations signal that there is no common agreement on the definition of what constitutes high quality provision.

5.3.1.2 Type 2: Postgraduate Research Ranking

Both Shanghai Jiao Tong University Academic *Ranking of World Universities* (SJTU *ARWU*) and the Times Higher Education (Quacquarelli Symonds Ltd) *World University Rankings* have exclusively focused on rankings in relationship to quality of research provision. The target audience could be considered to be potential funding sources and potential academic researchers. Typically, research ranking measures include articles published; papers cited; research student numbers, prestigious awards for research secured, etc.

The critique that follows applies to both types of ranking however the source and nature of the performance indicators used in the calculation will be substantially different, not only between the primary purpose of the ranking but also between the different publishers engaged in producing the rankings.

For example, a focus on the student market will focus on the nature of the learning environment, and facilities and resources available for the student, whereas a research focus will be on the track record for research secured by the university, and the research facilities available often coupled with peer esteem of the research status of the university.

The following section will draw, as appropriate, on both types of rankings.

5.3.2 Selecting Suitable Metrics: Performance Indicators

5.3.2.1 Performance Indicators

It is therefore not surprising that the Performance Indicator (PI) has helped form the landscape of higher education, providing a critical measure to help answer the question: How do I know what I am achieving? (Cave et al. 1997).

As such, performance indicators (PIs) are designed to provide quantifiable measurements which, having been agreed in advance, reflect the relative success of an organisation (Longden 2008). However, what is selected for measurement

is governed by the nature of the organisation and is political – political with a small '*p*'. Who decides to record student entry qualifications rather than student socio-economic background exposes a particular interest in the characteristics of students in higher education.

PIs are usually seen as numerical measurements of achievement that are easy to collect, interpret and use, with the emphasis on 'easy to collect'. In theory, PIs can only be derived from things over which direct control can be exerted leading to achieving an outcome of the measure. It is not surprising that PIs are of interest to a wide range of bodies, ranging from federal and local governments agencies, through to universities and colleges themselves, and, ever increasingly, students.

With the student market in mind, compilers of university rankings would claim that they have attempted to simplify a complex set of PIs measures by aggregating them to form a single index, sorted in order thus producing the university ranking.

The claim is made by compilers that university rankings 'help' potential students and their parents to reduce the mass of information about the universities and in doing so, they claim they are assisting in the decision making and enabling students to come to the conclusion about the 'right university to attend'.

With over 4,200 accredited universities in the USA and about 130 in the UK, for example, it is clear that the task facing a prospective student in selecting the 'right' university is a daunting one not only for the prospective students but also for concerned and interested parents. University rankings clearly service a need.

5.3.2.2 Proxy Measures

Given the origins of the data, it would not be surprising that compilers often require data that is not provided in the direct measurements provided by the sources discussed above. Teaching quality is one such measure that is deceptively simple and would be expected to be easily available but is neither. Compilers are forced to consider other ways of achieving the measure. In the UK, the measure is derived as a proxy measure from the *National Student Survey*[1] (NSS), while the *US News and World Report* in the USA derives the measure from a dubious logical connection between '*alumni giving and satisfaction*'. It could be argued that each PI should be scrutinised to ensure that what it measures and what it purports to represent in a ranking are sufficiently close to be acceptable.

5.3.2.3 From Judgement to Number: What Is Regarded as Important

Each event of observing the world evokes a judgment of what we decide to record about the event, and what particular part of the experience is important at that

[1]National Student Survey is conducted in the UK as a statutory requirement on all higher education providers to ensure that over 60% of their final year students contribute to the web based survey.

moment in time. Experience is not naturally coded as a set of numbers; we frequently impose a number at a later date and time when describing the event.

Within a commercial setting, it is possible to move from judgment to a numerical measure with greater ease than within the education setting, where it is often difficult, maybe impossible, to make hard measures from a socially constructed experience.

Graham and Thompson (2001) argue that most prospective students and their parents require reliable comparable information on the most important outcome of a college education, namely:

- What have I gained by way of learning from this experience–*learning outcomes*?
- Has the total experience rated highly on the *student satisfaction* index?
- Have I worked sufficiently effectively that I gained a certificate that will be acknowledged by others as a measure of success – *graduation*?

Interestingly, these apparently simple measures are dependent on proxy measures and rely on a simple relationship.

$$\text{Good student} + \text{Good faculty} = \text{Good university}$$

With this simple model, many compilers have set about to construct a university ranking that then teases out measures about faculty and students to help construct the metric of 'good university'. Frequently, compilers make use of measures relating to student entry qualifications, faculty qualification, i.e., percentage of doctoral staff, all of which are proxy metrics for measure that is more elusive to grasp- hold. Graham and Thompson suggest that:

> … [it is] like measuring the quality of a restaurant by calculating how much it paid for silverware and food; not completely useless, but pretty far from ideal. (Graham and Thompson 2001)

Despite best efforts, data are complied and reported according to value judgements that are embedded in the methodologies for collection and reporting; some of these value judgements are explicit, some implicit.

> Hardly a week goes by without another league table measuring university performance … . Of course none of these tables are the same; they all use different statistical measures and weightings to reach their judgments. While some focus on teaching quality, others emphasise research or take greater account of students' entry qualifications, the money spent by institutions, the proportion of students who get a 2:1 or the percentage who get a graduate job. Not only do these measures vary between papers, they also vary from year to year. So, while government teaching inspection scores might be important one year, it could be the level of library expenditure the following year. (Morris 2005)

5.3.2.4 Outputs, Inputs and Process

A helpful means of differentiating the different measures that are available in creating rankings, be it for ranking universities or subject within universities or research generated by universities, is to classify the measures into the three types of PI – outputs, inputs and processes.

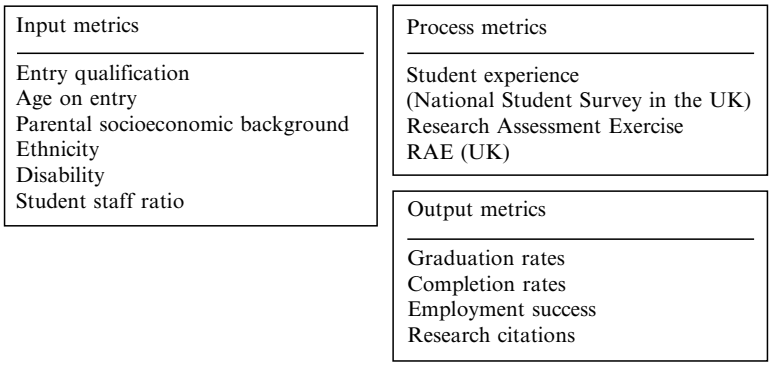

Fig. 5.1 Broad categories and types of metric

Input measures might include qualifications on entry, student staff ratios, resource expenditures, while output measures might include successful completion rates, employment characteristics and degree success rates, citations or published articles. Processes, while being harder to measure, relate to the quality measures for research or teaching; for example, in the UK, the National Student Survey (shown as a process measure in Fig. 5.1) provides a measure of the student experience; it does not provide a measure of the student learning.

The point is well made by Richardson (2008: 18) that few compilers in the UK make any clear distinction between the three types of metrics shown in Fig. 5.1 and that the over-emphasis on input is to the detriment of the overall ranking methodology. Richardson, citing work by Pascarella and Terenzini in 2005, suggests that there is little evidence to support input measures as they 'have only an inconsequential or trivial relationship with student learning and achievement'.

5.3.3 Collecting Data

Both UK and US compilers attempt to make maximum use of authoritative data. As will be discussed later in this chapter, confidence in how authoritative the data may be has been challenged in the USA. In the UK, the data used by most compilers has been collected through an agency of central government.

The creation of a ranking index depends on the selection of data, data originating, as I have argued, from events of different forms and complexity within the life of the university. Compilers of university ranking indices draw on this data to help construct the ranking. Three types of data are available for use in the ranking tables,

- Primary data generated by the university itself
- Survey data generated by the compilers
- Data collected from independent third parties

Primary data produced by universities for both internal and external purposes has been subjected to external audit. In the UK, there is a statutory responsibility placed on all publicly funded institutions to provide data to Government Agencies to support the financial investment made. The data is subject to tight auditing which requires clear precise definitions of the data used, specific dates around which data is collected, recorded, and transmitted. It is this audited data which, if released to the commercial compilers, is used to create the ranking index. In the USA, data provided to the commercial compilers is provided directly by the institution. The audit function is absent. Usher and Savino note that in the USA:

> there is no guarantee that institutions will actually report the data to the rankers on a consistent basis , as all have a clear incentive to manipulate data in a manner which will benefit them (Usher and Savino 2007: 26)

This prophetic insight by Usher and Savino has materialised itself in revelations from two colleges in the USA. In both cases, the risk implicit in self-regulation associated with submitting unaudited data to the compilers of ranking tables is evident.

The first case is that of Clemson University, which claims to be one of America's top public research universities, located in South Carolina. A conference presentation[2] by a member of Clemson University staff exposed the practice of managing data to secure an improvement year-on-year on the rankings. Among the steps reportedly alleged by Watt, who until 2006 headed Clemson's institutional research office were that Clemson:

- Manipulated class sizes
- Artificially boosted faculty salary data
- Gave rival schools low grades, which counts for 25% of the score in *US News* and *World Report's* peer reputation survey

Irrespective of the actual truth in the claim and counter claim, the weakness is there for all to see. In essence, Clemson University submitted data that ensured the University rankings moved from 38th to 22nd position in *U.S. News*' ranking of public research universities from 2001 to 2008.

> The easiest moves … revolved around class size: Clemson has significantly increased the proportion of its classes with fewer than 20 students, one key U.S. News indicator of a strong student experience. [I]t has focused … on trying to bump sections with 20 and 5 students down to 18 or 19, but letting a class with 55 rise to 70. 'Two or three students here and there, what a difference it can make', Watts [Clemson administrator responsible for managing the US News submission] said. It's manipulation around the edges. (Lederman 2009a)

By creatively managing the class size data in this way, it was possible to ensure that Clemson University PI for student data was maximised for the faculty resources element of the index.

[2]Title of AIR presentation Strategic Project Management: Lessons from a Research University's Successful Drive to Move Up in the US News Rankings by Catherine E. Watt – Director of the Alliance for Research on Higher Education, and Nancy T James – Research Analyst III, Clemson University.

How widespread an occurrence of this 'creative management' of PIs is hard to assess – that it occurs at all is no longer in doubt.

The second case relates to self reporting of data submitted to *US News* and *World Report* by the University of Southern California (USC). USC claimed that 30 of its professors were members of the prestigious National Academy of Engineering; on its Web site, the engineering school went even further by listing 34 such professors (Lederman 2009b; Shea 2009). Further investigation provided evidence that the claim was a substantial over estimate of the actual number of 22.

Clemson and USC are not the only institutions susceptible to the pressures to managing their ranking status. Ehrenberg (2003) in the USA and Watson (2008) in the UK have written on this topic.

Survey data may be developed specifically by the compilers (as in the case of *US News* and *World Report* which incorporates a peer reviewed surveys (see Table 5.3) contributing 25% to the final ranking score) or may be extracts from national surveys as in the case of the UK's *National Student Survey* (NSS).

> The National Student Survey, which measures student satisfaction, will be factored into the rankings for the first time. As a result, figures that represent the subjective sentiments of those who are willing to fill in the forms will be turned into seemingly objective measures of the worth of higher education institutions (Ryan 2009).

The NSS data was developed to provide a measure of the quality of the student experience of higher education in the UK. It remains a statutory responsibility for each higher education institution in receipt of public funds to provide a minimum survey response rate[3] anonymously completed by its final year students. While there have been claims of influencing the student opinion about their experience (Newman 2008), the extent is limited.

The use of surveys developed by compilers remains an important component of the *US News* and *World Report's* methodology. The limitations of this approach have been highlighted by the Clemson clarification that the President's completion of the Peer review submission exposes the difficulties in being altruistic when self reporting.

> … Clemson officials, in filling out the reputational survey form for presidents, rate "all programs other than Clemson below average," to make the university look better. "And I'm confident my president is not the only one who does that" (Lederman 2009a)

The Times rankings in 2008 introduced the approach well established by *US News* and *World Report* by seeking the opinions of Heads of secondary schools and from university academics about where the highest-quality undergraduate provision was located. It was probable that *The Times* had borrowed the approach from *US News* and *World Report* – wisely, this type of data collection was dropped for the following year calculations!

Independent third party data, usually from administrative source such as government or grant making bodies, are generally regarded as '*the gold standard of*

[3]NSS – the minimum response rate for 2009 was 60%.

Table 5.3 US News and World Report Weights for national universities and liberal arts colleges

Metric		Sub-factors		Overall contribution – weighting	
		National universities and liberal arts colleges (%)	Universities master's and baccalaureate colleges (%)	National universities and liberal arts colleges (%)	Universities master's and baccalaureate colleges (%)
Peer assessment	Peer assessment survey	100	100	25	25
Student selectivity	Acceptance rate	10	10	15	15
	High school class top 10%	40	0		
	High school class top 25%	0	40		
	SAT/ACT scores	50	50		
Faculty resources	Faculty salary	35	35	20	20
	Staff qualifications	15	15		
	Full-time faculty	5	5		
	Student faculty ratio	5	5		
	Class size 1–19 students	30	30		
	Class size 50+ students	10	10		
Graduation and retention rate	Average graduation rate	80	80	20	25
	Average freshman retention	20	20		
Financial resources		100	100	10	10
Alumni giving	Average alumni giving	100	100	5	5
Graduation rate	Graduation rate performance	100	0	5	0

Source: Morse and Flanigan (2009)

comparative data since they are, at least theoretically, both accurate and impartial' (Usher and Savino 2007: 26). In their survey of global ranking systems, Usher and Savino make the point that while accuracy and impartiality might be positive aspects of their contribution, they are really administrative by-products of data collected for other purposes, opening up the potential for using a measure out of its original context.

The plurality in use of data sources varies considerably even within one country where two or more commercial compilers operate. This should raise concerns among those who depend on the rankings as it implies inconsistency in the principles adopted by each complier as to where the data comes from.

The '*Times* ' (Times online 2009), for example, describes the 2010 tables data sources as:

> All sources of the raw data used in the table are in the public domain. The National Student Survey (NSS) was the source of the Student Satisfaction data. … The information regarding Research Quality was sourced from the 2008 Research Assessment Exercise … Staffing data supplied by HESA were also used to evaluate the extent to which the research ratings related to total academic staff. … Entry Standards, Student-Staff Ratios, Services & Facilities Spend, Completion, Good Honours and Graduate Prospects data were supplied by the Higher Education Statistics Agency (HESA) which provides a system of data collection, analysis, and dissemination in relation to higher education in the whole of the United Kingdom. The original sources of data for these measures are data returns made by the universities themselves to HESA (Times online 2009).

The Guardian (MacLeod and Hiely-Rayner 2009b), drawing on similar sources but from a very different set of performance indicators, have used the HESA and NSS data in their 2009 calculation of a ranking metric.

> The rankings are compiled from the most recent figures available - official 2006-07 returns from universities and higher education colleges to the Higher Education Statistics Agency (Hesa). They also incorporate data from the National Student Survey (NSS) 2007, published by the Higher Education Funding Council for England (Macleod and Hiely-Rayner 2009a).

The pretence at being objective and quasi scientific has been discussed by Stella and Woodhouse (2006: 6), where they suggest that there are generally two broad data types: data provided by institutions and data derived from expert opinion, both giving an illusion of being 'scientific' and by inference, beyond reproach.

> Most rankings rely on two types of data - data given by institutions that is accepted, often without a reliable validation process, and data obtained from opinion polls in the name of 'expert opinion'. With both components on shaky grounds, the use by the media groups of complex formulae with weights and indicators only helps to project a pseudo image of being 'scientific' to outcomes that may be statistically irrelevant (Stella and Woodhouse 2006: 10).

There is a clear necessity for data to be managed within an institution in ways that take account of the uses to which they are, and might be, put. The importance of data definition and management for how the data is returned to the compilers of the rankings is such that, at all levels within an institution, staff are aware of the ways in which what they do, and how it is recorded, could have a significant impact on their futures.

A sudden decline in an institution's position in the rankings, which might derive simply from particular choices in collating and reporting data, could for instance have a sharp adverse effect in the international market for higher education. It matters greatly how an institution presents truths.

The process of creating a university ranking index starts within the university and the events that constitute that university. These events are various, complex, and frequently invisible to the casual viewer; some of the events are captured, nonetheless, for different reasons by the university. The reasons can be various too; faculty management requirements such as class lists, assignment submissions etc.; internal management of the university to ensure quality standards are maintained or facilities are supported; and external statutory requirements such as those required in the UK to support the funding model used by the Higher Education Funding Council for England (HEFCE) to distribute block teaching funds. The list of events and therefore data is substantial. From this mass of data, compilers select certain items of data to include in the ranking methodology.

The apparent simple act of capturing data brings with it intrinsic difficulties. To a university outsider, the simple event of counting the number of students on a program would not appear too challenging. However, those involved in data collection are only too well aware that data collection brings with it a set of ever expanding definitions. The quotation from HESA (2009) illustrates the increasing complexity of the data definition required by universities. Data submitted in the UK is provided to both the Higher Education Funding Council for England (HEFCE) – which is an aggregate data set – and to the Higher Education Statistics Agency (HESA), where an individual's student record is submitted. Considerable pressure is placed on universities to ensure that data quality is high; both HESA and HEFCE have sophisticated data audit systems operating to ensure consistent, accurate data is provided. As a final pressure on universities, HEFCE operate a data audit on institutions to maximise data quality.

Subsets of the data are released by HESA, after data protection agreements for each data request has been agreed (or not as the case may be), to UK compilers of university rankings.

It is important to note at this point that the data provided in the England to HESA was originally provided to support funding claims against HEFCE; the data was not collected to assist compilers with the production of their university rankings.

> The HESA session population has been derived from the HESA Student Record. It includes all higher education and further education student instances active at a reporting institution at any point in the reporting period 1 August to 31 July except: dormant students (those who have ceased studying but have not formally de-registered) incoming visiting and exchange students. Students where the whole of the programme of study is outside of the UK, and from 2007/08: students on sabbatical.
>
> Incoming visiting and exchange students are excluded from the session population in order to avoid an element of double-counting with both outgoing and incoming students being included. The HESA session population forms the basis for counts of full-time equivalent (FTE) student instances (HESA 2009).

In the extract above from the HESA guide to higher education institutions for submission of data in 2009, the complexity and need for very precise definitions

is evident. This is partly why comparative data is difficult to obtain. Where can there be confidence in the precise mapping of data across educational jurisdictions? What, for example, is implied by the deceptively simple term "de registration"?

5.3.4 Adaptation of PIs into a Scale

The selection of metrics for inclusion reflects the objectives that are to be achieved by the ranking process. In *US News*, *The Guardian,* and *Times,* a measure used in the overall ranking calculation is a financial one. For example, faculty compensation in *US News* is the average faculty pay and benefits adjusted for regional differences in cost of living. In the case of *The Guardian* and *Times,* the data is taken directly from the HESA finance return and is a ratio of spend per student full time equivalent (FTE). In both examples from the USA or UK, the final measure is $ per faculty or £ per student FTE. It would not be possible to incorporate these values directly into any calculation of ranking without an adaption.

Richardson (2008: 20) notes that the process of adapting the data, in readiness, for aggregation is frequently termed, incorrectly (in the strict statistical sense), as normalisation. It encompasses the process of adapting the data to reflect adjustments necessary when dealing with institutional size or institutional subject/discipline composition. It is acknowledged in the UK that the national funding model positively advantages institutions with significant medical schools when spend per student is considered (evidence from the USA and Australia suggest a similar effect occurs there too). Compilers, in their attempt to deal with this distortion, apply a modification to the metric to account for this 'distortion'.

Data used by the Guardian's 2009 guide for spend-per-student studying. Sociology indicates that the range of data is from £407.99 to £3,243.49. This is calculated from the amount of money that an institution spends providing a subject (not including the costs of academic staff, since these are already counted in the staff-student ratio) adjusted to account for the variation in volume of students learning the subject. Figure 5.2 also includes the money the institution spends on central academic services, and per student FTE.

In discussion with a compiler of the Guardian's table, it became clear that while the actual data was incorporated directly into the calculation, for display and publication purposes, and to avoid issues related to publishing actual data in the table, the data was transformed to a single 1–10 scale.

The adaptation of data into a scale is frequently used in the methodology adopted by *The Times*. The construction of the scale is arbitrary and not based on any theoretical analysis. The assumption is that the scale is linear; but there is no justification for that assumption. Why not log linear or inverse or sigmoid?

Either the ranking lends itself to a scale of 0–100 or to a band to which numerical values can be applied. Whichever detailed process is used, the final product is a numerical value for the PI which can then be used directly in producing the final index.

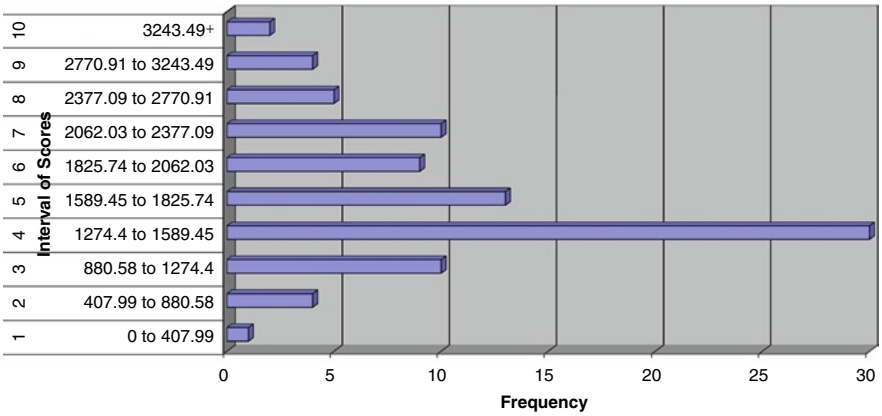

Fig. 5.2 Distribution of spend per student in sociology (Source: MacLeod and Hiely-Rayner, 2007)

It is interesting to note that the instructions provided by *US News* on the Faculty Compensation measure (see Morse 2009 for a detailed definition of the measure) hint at the importance of the metric in the overall calculation of the ranking score. It notes that "…*higher average faculty salaries score better in the ranking model*" (Morse 2009). The model used by *US News and World Report* in the construction of their college ranking tables rewards institutional expenditure over any other institutional measure used in the calculation of the final ranking score.

5.3.5 Standardising Measures Prior to Aggregating

A relatively inconsequential paper published in 1955 provides a challenge to anyone attempting to add different types and sources of data together (Richmond 1968: 182). The paper published in the *Journal of incorporated Associations of Assistant Masters in Secondary Schools* describes a simple scenario where test scores for ten subjects were set out in a table for ten children. The scores were added together to provide an aggregate score for each individual. From the total, it becomes clear who should be labelled '*top of the class*'. However, on closer scrutiny it becomes clear that each subject has used a different range of marks – some used the whole scale from 0 through 100, while others use a scale from 30 through 65. When the variability in use of the scale is incorporated into the calculation, the rank order in the class is reversed.

The implications for those engaged in aggregating data from different sources and from different distributions are both significant and important.

Adding scores together is simple but it can only be acceptable when the scores have been adjusted so that the distribution and range of the scores conforms to a specific format. In the next section, consideration is given to the nature of that format and the conditions that need to be achieved before confidence can be given to the outcome.

5.3.5.1 Transformations

To ensure that when adding the two data sets together the same 'measures' are added together, it is necessary to treat the two data sets by stretching them so that they conform to common statistical measures where the mean value is zero and the standard deviation is 1. This transformed distribution thus created is referred to as the 'z-score'. The 'z-score' transformation requires some basic understanding of statistics.

The problem of adding two PI scores is best illustrated in Table 5.4 based on data from 26 universities. A closer statistical examination of each of the PIs shown in Table 5.4 indicates that the range of data for *PI* A ranges from 90 to 20 while that for *PI* B ranges from 62 to 43. When the mean scores are compared, a further

Table 5.4 University rank order created from two performance indictors

University	*PI* A	*PI* B	Sum (*PI* A + *PI* B)	Rank order
U_001	*85*	*54*	139	**1**
U_002	*85*	*50*	135	**2**
U_003	*90*	*44*	134	**3**
U_004	*74*	*51*	125	**4**
U_005	*78*	*46*	124	**5**
U_006	*76*	*44*	120	**6**
U_007	*70*	*50*	120	**6**
U_008	*64*	*53*	117	**8**
U_009	*62*	*55*	117	**8**
U_010	*64*	*52*	116	**10**
U_011	*60*	*56*	116	**10**
U_012	*62*	*52*	114	**12**
U_013	*64*	*45*	109	**13**
U_014	*45*	*61*	106	**14**
U_015	*51*	*54*	105	**15**
U_016	*47*	*57*	104	**16**
U_017	*50*	*51*	101	**17**
U_018	*54*	*47*	101	**17**
U_019	*34*	*62*	96	**19**
U_020	*28*	*57*	85	**20**
U_021	*42*	*43*	85	**20**
U_022	*35*	*50*	85	**20**
U_023	*30*	*54*	84	**23**
U_024	*35*	*49*	84	**23**
U_025	*24*	*50*	74	**25**
U_026	*20*	*44*	64	**26**
Min	*20*	*43*		
Max	*90*	*62*		
Sum	*1429*	*1331*		
Mean	*54.96*	*51.19*		
Number	*26*	*26*		

difference becomes evident. (*PI* A = 54.96 and *PI* B = 51.19). This illustrates clearly that the two data sets are different and that any attempt at aggregating each of the individual PIs together would present a problem.

5.3.5.2 *z*-Scores: Calculating Standardised PIs

The '*z-score*' provides two important characteristics about performance indicators

- The relative position of the PI measure relative to the mean
- The distance from the mean

Negative 'z-scores' indicate PI measures below the mean; positive *z*-scores indicate PI measures above the mean. 'z-scores' with a larger absolute value are further away from the mean from *z*-scores that those that are smaller in absolute value (−2.30 is further from the mean than .40).

The calculation of a 'z-score' can be considered in two stages.

The first stage introduces the concept of spread of data around the mean. The mean value ($\overline{x}$ described as bar x) is calculated from a summation of the all the university scores divided by the number of scores contributing to the total score. The spread of data for each university from the mean is aggregated, i.e., $x - \overline{x}$, where x is the individual value for each university and $\overline{x}$ is the mean for all the universities.

$$z = \frac{x - \overline{x}}{SD}$$

From Table 5.4, the mean value for all universities for *PI* A is 54.96; the value specifically for university 004 is 74, giving a difference from the mean of 19.04 (when each individual measure from the *PI* A mean is aggregated the net result is a mean of the spread or variance value of zero).

The second phase of standardisation involves a calculation of the standard deviation of the university PIs. To achieve this, each variance score is squared and aggregated and then divided by the number of universities contributing. The formula is shown below:

$$s = \sqrt{\frac{r_1^2 + r_2^2 + \cdots + r_n^2}{n-1}} = \sqrt{\frac{\sum \left(x_i - \overline{x}\right)^2}{n-1}}$$

When this is carried out for each university score both positive and negative numbers result. Statisticians frequently use the device of squaring a number to remove the problem of negative values. The standard deviation for *PI* A is 20.10.

The amended Table 5.5 (based on Table 5.4) has been extended to show the calculations of the *z-score*s for each university together with a comparison of the difference between the rank based on the raw data from the ranking derived *z-score transformations*.

Table 5.5 Comparison between rank order created from raw scores and transformation score ('*z* score')

	RAW				*z* score				
University	*PI* A	*PI* B	Sum (*PI* A + *PI* B)	Rank	*PI* A	*PI* B	Sum *z*(*PI* A + *PI* B)	Rank	Differences
U_001	85	54	139	**1**	1.49	0.11	1.60	**1**	***0***
U_002	85	50	135	**2**	1.49	–0.05	1.45	**3**	***–1***
U_003	90	44	134	**3**	1.74	–0.28	1.47	**2**	***1***
U_004	74	51	125	**4**	0.95	–0.01	0.94	**5**	***–1***
U_005	78	46	124	**5**	1.15	–0.20	0.95	**4**	***1***
U_006	76	44	120	**6**	1.05	–0.28	0.77	**6**	***0***
U_007	70	50	120	**6**	0.75	–0.05	0.70	**7**	***–1***
U_008	64	53	117	**8**	0.45	0.07	0.52	**8**	***0***
U_009	62	55	117	**8**	0.35	0.15	0.50	**9**	***–1***
U_010	64	52	116	**10**	0.45	0.03	0.48	**10**	***0***
U_011	60	56	116	**10**	0.25	0.18	0.44	**11**	***–1***
U_012	62	52	114	**12**	0.35	0.03	0.38	**12**	***0***
U_013	64	45	109	**13**	0.45	–0.24	0.21	**13**	***0***
U_014	45	61	106	**14**	–0.50	0.38	–0.12	**15**	***–1***
U_015	51	54	105	**15**	–0.20	0.11	–0.09	**14**	***1***
U_016	47	57	104	**16**	–0.40	0.22	–0.17	**16**	***0***
U_017	50	51	101	**17**	–0.25	–0.01	–0.25	**18**	***–1***
U_018	54	47	101	**17**	–0.05	–0.16	–0.21	**17**	***0***
U_019	34	62	96	**19**	–1.04	0.42	–0.63	**19**	***0***
U_020	28	57	85	**20**	–1.34	0.22	–1.12	**23**	***–3***
U_021	42	43	85	**20**	–0.64	–0.32	–0.96	**20**	***0***
U_022	35	50	85	**20**	–0.99	–0.05	–1.04	**21**	***–1***
U_023	30	54	84	**23**	–1.24	0.11	–1.13	**24**	***–1***
U_024	35	49	84	**23**	–0.99	–0.08	–1.08	**22**	***1***
U_025	24	50	74	**25**	–1.54	–0.05	–1.59	**25**	***0***
U_026	20	44	64	**26**	–1.74	–0.28	–2.02	**26**	***0***
Sum	1429	1331			0.00	0.00			
Mean	54.96	51.9			0.00	0.00			
Number	26	26			26	26			
SD	20.10	5.15							

5.3.5.3 Impact

To illustrate the impact of the two ranking methodologies, Fig. 5.3 provides a graphic for the difference between the raw ranking position and the position based on the 'z-score' transformation. For some universities, the impact is insignificant; for example, U_001 is un-affected by the transformation and remains at the top of the rankings, whereas U_020 based on the raw score ranking was 20th, yet when based on the 'z-score' transformation was adjusted to the 23rd position. The impact of applying a 'z-score' transformation to the raw data shows up very clearly in the apparent random changes that occur between the universities.

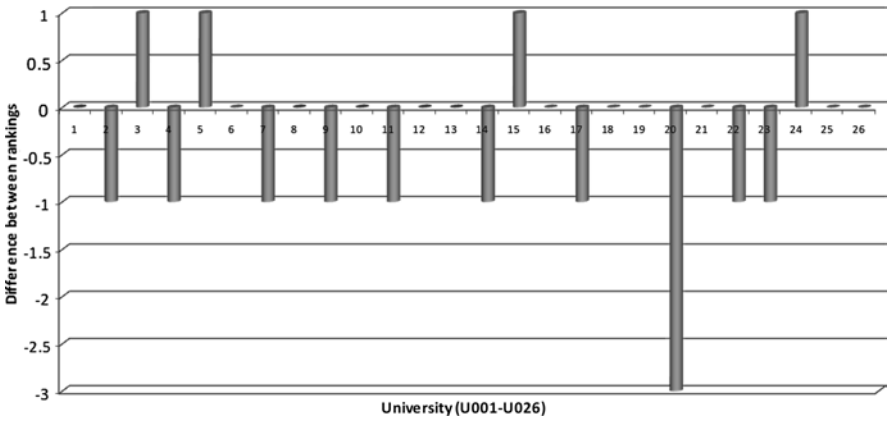

Fig. 5.3 Impact of applying a transformational (*z* score) when aggregating two PIs

In summary, the '*z*-score' provides a measure of the number of standard deviations (SD) each *PI* measure is away from the mean. For example, a '*z*-score' of 1.3 means that the PI was 1.3 SDs above the mean, whereas a *z*-score of −.70 means that the PI is .70 SDs below the mean and a *z*-score of 0.00 indicates a PI exactly the same as the mean.

By re-calculating *z*-scores for each PI, we have essentially re-scaled, or re-numbered the scores. In other words, we have essentially changed the scores from their original values to new values that are directly interpretable. Because *z*-scores are linear transformations, we have not changed the shape of the distribution.

For a detailed explanation of the underlying theory associated with '*z* score' transformations it is suggested that you refer to appendix A of Richardson's report (2008: 6). Standardisation refers to the process of mapping a set of performance measures onto a single scale where the standard deviation is one and the mean value is zero (see Hinton 2004; Miles and Shevlin 2006 for detailed accounts of the underlying mathematics).

This again raises important questions such as "Is the process of standardisation incorporated and applied by the compilers in preparation of their rankings though?"

5.3.6 Weighting PIs Prior to Aggregating

Anyone who has had the experience of adding oil to petrol to run a two stoke petrol engine knows exactly what ensuring the correct mix means. If the mix is 25:1, then making the mix 15:1 will result in trouble! In this example, there is a theoretical underpinning, beyond my understanding admittedly, that justifies why 25:1 is the correct mix. How does this relate to university rankings?

In the creation of a single ranking index, several measures are added together. But what is the mix or the weighting applied. It is not difficult to realise that given

Table 5.6 Guardian PI measures and relative weightings used creating the 2010 rankings

We have used seven statistical measures to contribute to the ranking of a university or college in each subject, weighted as follows:

- Teaching quality: as rated by final year students on the course (10%)
- Feedback (assessment): as rated by final year students on the course (5%)
- Spending per student (17%)
- Staff/student ratio (17%)
- Value added: comparing students' degree results with their entry qualifications (17%)
- Entry score (17%)

Source: MacLeod and Hiely-Rayner (2009b)

Table 5.7 Maclean's 2008 weighting and measure

Category	Sub-factor	Weighting (%)
Students/classes	Student awards	20
	Student faculty ratio	
Faculty	Faculty awards	18
	Faculty grants	
Resources	Research income	12
	Operating budget	
Student services	Scholarships and bursaries as % of budget	13
	Student services as % of budget	
Library	Expenditure	15
	Requisitions	
	Holdings per student	
Reputation	Survey	22

Source: Dwyer (2008)

so many global university ranking tables, there is no agreement of the relative contribution of the measures. In simple terms, the weightings adopted by compilers are idiosyncratic and devoid of a theoretical underpinning.

Looking more closely at Table 5.6, what the *Guardian 2010* compilers are implying is that qualifications on entry (17%) are more important in the contribution to the overall ranking index than the teaching quality which contributed 10%. On what basis are these 2% based? Who decides that one measure contributes more to the overall measure of the university?

To show that this is not restricted to the UK, consider the way in which Maclean's university ranking operates for students in Canada. Maclean's, unlike compilers in the UK, places universities in one of three categories, recognising the differences in types of institutions, levels of research funding, the diversity of offerings, and the range of graduate and professional programs. The three categories are primarily: Undergraduate universities where few graduate programs are available; Comprehensive category where there is a significant volume of research and there are many graduate programs on offer; and finally those defined as Medical Doctoral Universities where a broad range of Ph.D. programs and research are provided and where there are medical schools, which set them apart in terms of the size of research grants (Table 5.7).

It might be expected that an analysis of the PIs, that the compilers plan to use, might be statistically analysed to tease out the relative importance of the contributing PIs. It would be possible to use Factor Analysis or Logistic Regression to secure a measure of confidence in the relative importance of the PIs to each other and therefore to the final ranking index.

Why is it the case that compilers do not follow this path?

Combining data may appear innocuous but little research has been conducted that allows us to make a balanced judgment as to the balance that ought to be applied. The principle involved in the process of weightings involves assigning to each indicator a weight that reflects the perceived importance and then combining these weights into an overall score.

> Nonetheless, just as democracy, according to Winston Churchill, is the worst form of government, except for all the others, so quality rankings are the worst device for comparing the quality of … colleges and universities, except for all the others. (Webster 1986)

Shapiro (1994), principal and vice-chancellor of McGill University at the time, commented on the shortcomings of *Maclean's* ranking publication and drew attention to graduation rates. A university with high graduation rates could either be a university 'providing effective education and support to excellent students or a university with lax evaluation and standards'. His letter to the editor of *Maclean's* questions the logic of combining indicators to:

> …obtain a global evaluation or ranking is the most difficult for Maclean's to rationalise. The process requires a decision on the weight to be assigned to each parameter in the equation. These weights must arise from value judgements on which there will never be universal agreement. In Maclean's case, these are based on the values of the Maclean's editors. It is quite clear that a different set of values could result in a different global evaluation and ranking…and it is impossible to determine objectively which set of values and weights is to be preferred (Shapiro 1994).

The compilers start with a mass of data and through a series of mathematical and statistical procedures reduce the data to a single column. In the example provided above relating to how standardisation of scores can assist in the process of aggregating two PIs, the implicit assumption was that the two PIs would be aggregated like for like. The assumption had no theoretical foundation why should *PI* A contribute equally with *PI* B to the overall score.

5.3.7 Creating a Single Index Reflecting a University

The primary objective of the university, subject, or research ranking is to end up with a single measure – a metric – that constitutes a measure of quality. The final stage is then to sort the indices into an order from high to low. No account is taken at this final stage of the significance of any differences between the indices that emerge from the processes described above Richardson (2008: 14).

5.3.7.1 Statistical Difference

Gerhard Casper, then president of Stanford University, in a letter of protest to the editor of the *US News and World Report*:

> "…Could you not do away with ranks ordering and overall scores, thus admitting that the difference between #1 and #2 – indeed between #1 and #'10 – may be statistically insignificant." (Casper 1997)

Clarke, citing a more extensive quotation from this letter, raises the question that while the issue has received much debate but acknowledges that little research has been conducted on the implications.

The significance of difference in scores is not easy to judge from a ranking table where small differences in scores can translate to substantial differences in ranking because of heavy clustering around the mean. The *Times Higher World Ranking of Universities* (QS 2009) in the subject cluster Social science finds a difference of just 1.2 points on a 100 point scale between rank 83 and 89. In the overall university rankings, there is just a 1.9 point difference between rank 64 and 73 going down to a slim 1.2 point difference between rank 65 and 72.

5.3.7.2 Volatile Rankings

Confidence with the constancy of university rankings may be challenged by the data that follows. The data represented in Fig. 5.4 is taken from rankings created by the Guardian Newspaper in June 2009. The graph shows on the left hand scale the rank index order for the UK[4] universities based on the order for 2010. Universities are represented by the column that increase from left to right (light shading); superimposed on that graphic is a secondary graph that reports the difference between the 2010 and 2009 ranking position for each university. It is possible to suggest that:

- Small fluctuations in the size and number of dark bars (indicative of changes between the two years) are indicative of ranking consistent across years. Little change occurs and the rank order is resilient.
- Large fluctuations in both the number and size of the dark bars (indicative of change between the two years) are indicative of turbulence (Longden and Yorke 2009) between the years.

What are the implications of such a volatile system? Is it possible for an institution to change its relative position to other institutions from 1 year to the next? Figure 5.4 shows that fluctuations, or turbulence, occurs randomly among institutions from year to year. The graphic shows 2008 university ranking for UK universities in light grey with dark bars superimposed on the base data for 2008, representing the change between 2008 and 2009 ranking data.

[4]In 2009, five higher education institutions refused to release data held by HESA to compilers involved in the creation of university rankings. There is evidence that this is an increasing trend in the UK. In 1999, there was a single institution refusing to release data.

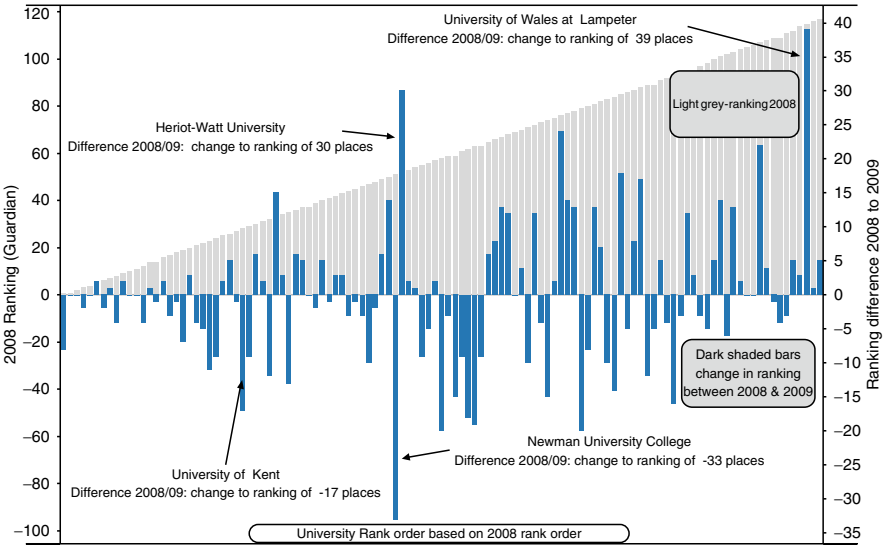

Fig. 5.4 Change in ranking index for UK universities between 2008 and 2009 (Source: MacLeod and Hiely-Rayner, 2009b)

Measurements recorded above the zero base line arise from positive movements in 2009 compared to the base line data for 2008. Measures below the zero base line indicate institutions where the 2009 placement is below that for 2008.

If the system were stable, then there should be few if any dark bars superimposed as the difference would be zero. Interpreting the data in Fig. 5.4 suggests that:

- A large number of dark bars (difference between 2008 and 2009) superimposed over the 2008 ranking, implying that many changes in university rankings occur between the years.
- The length of the dark bars (difference between 2008 and 2009) provides a visual indication of the size of the difference; for example, The University of Wales at Lampeter moved 39 places.
- The dark bars indicate that some of the differences are substantial, both positive and negative, implying that for some universities, the change in ranking is significant.

There are several explanations that can be considered to account for this. It could be related to changes in the methodology between the years in question giving rise to the fluctuation. It could also be related to internal institutional behaviour. The behaviour of Clemson could be considered to cause such a fluctuation, but it might also relate to negative outcome from internal reorganisation.

This raises further important questions that need to be addressed if confidence is to be restored. Should readers be informed about the volatility of rankings and that an institutional ranking may be subject to wide variation between the year the data was collected and the student formally engaging with the institution?

When data prepared by QS for The *Times Higher World Ranking of Universities* (QS 2009) is subject to the same analysis the results for 2007 and 2008 are shown in

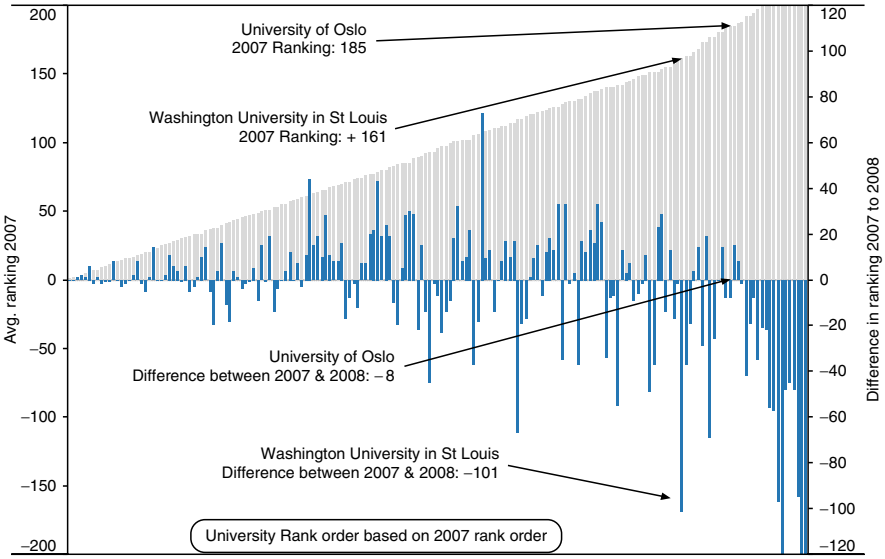

Fig. 5.5 World ranking volatility between 2007 and 2008

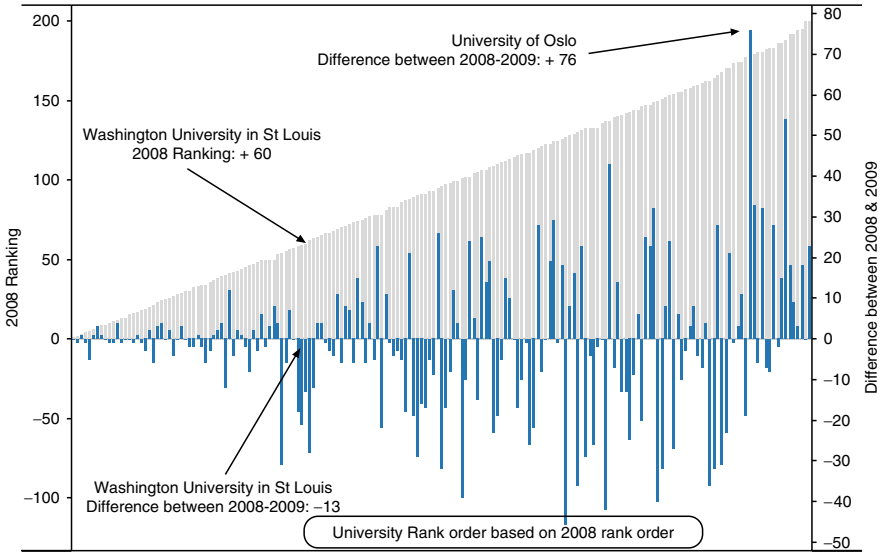

Fig. 5.6 World ranking volatility between 2008 and 2009

Fig. 5.5 and for data relating to 2008 and 2009 in Fig. 5.6. The graphics show clearly that considerable turbulence is evident over the two years. The turbulence appears to be greater in Fig. 5.5 towards the lower rank in the order of universities, although both graphics suggest extensive and substantial variation in differences between two years of data. For example, Washington University in St Louis from 2007 to 2008 moved

down 101 places but moved up 60 places by the time 2009 data was published. A similar pattern can be detected for the University of Oslo, which dropped 8 places between 2007 and 2008 but moved up 76 places by the end of 2009.

The three graphics (Figs. 5.5–5.7) provides evidence that the turbulence is not limited to teaching or research but is evident in both forms of university activity.

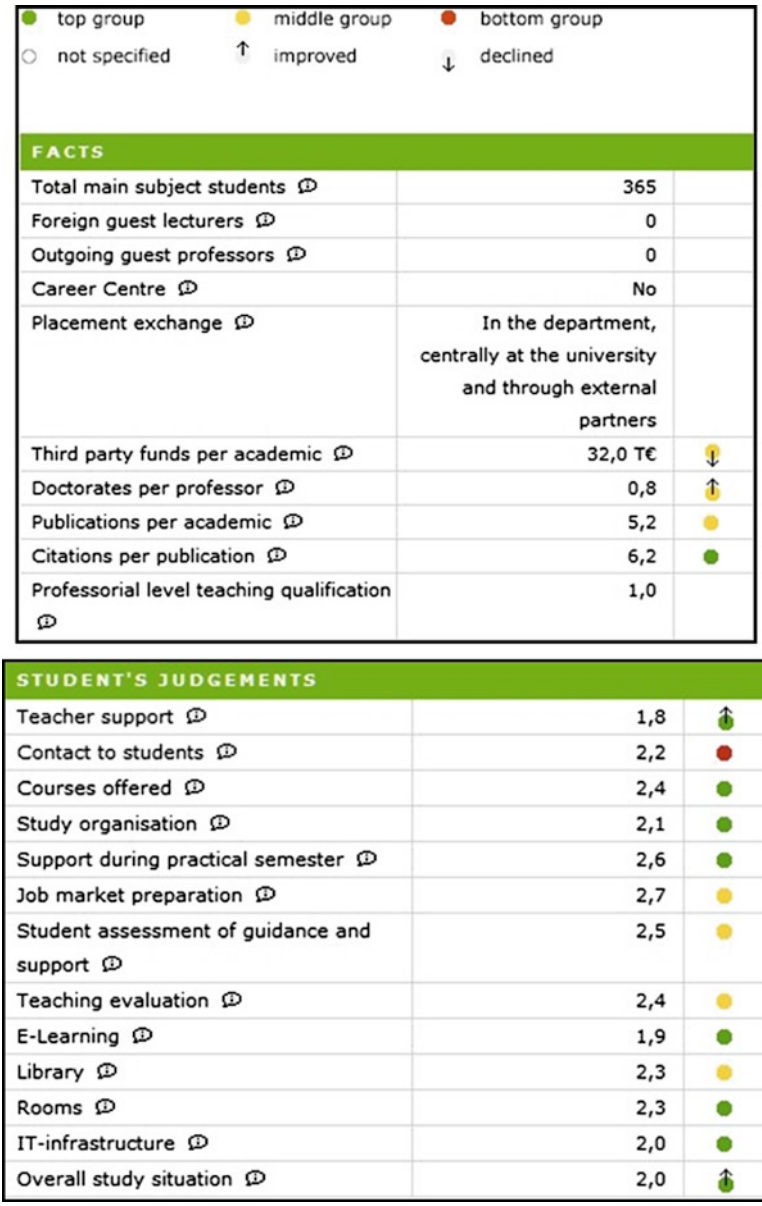

Fig. 5.7 Typical CHE university ranking data for subject 'X' at university 'Y' illustrating the 'traffic light' presentation of ranking data

5.3.7.3 Reality Check

Anecdotal conversations are frequently cited where compilers when asked what they would do, if the final ranking indices were unexpected and elevated an institution substantially beyond the anticipated or expected position, suggest that they would adjust the algorithm. It is often referred to as the reality check. It raises the question about '*whose reality is being used as the benchmark for checking*?' It may even suggest that the algorithm is derived from the expected ranking rather than the other way round!

It also raises the ethical question for the compilers when an unexpected index occurs, 'what should they do next!'

5.4 Positive Developments

The focus in this chapter has been on methodology drawing on examples drawn mainly from rankings published in the UK, USA and Canada. Increasingly, most publishers are making greater use of the potential that publishing on the web can offer. A recent development made possible through the web is the development of an interactive approach, leaving it to the reader to select key indicators in the creation of an overall score. This approach does not vindicate the criticisms discussed in this chapter, but quite the reverse, because it passes responsibility for measuring 'quality' from the publisher to the potential student.

An innovative approach has been developed at the Centre for Higher Education Development (2009) in Germany, designed to address the needs of providing quality information to prospective first-year students and the need to identify research performance quality.

The methodology used for University Ranking (CHE-Hochschul Ranking) relies on data relating to the departmental/subject level in contrast with the usual interest of ranking at the institutional level. By making this decision, CHE-Hochschul Ranking acknowledges that many weaker institutions have national or world class departments that would otherwise be overlooked. It also rejects the concept of 'best HEI'.

At the heart of the methodology is the idea that universities and colleges have individual strengths and weaknesses and that there are no empirical or theoretical bases on which to give weighting to individual factors. It argues that, as the target group is first year students where they are heterogeneous in their preferences, it would be counterproductive to use fixed predetermined weightings.

Instead the HEI is 'viewed' from several different perspectives – professors, managers, students. Each allows for contrast to be made between subjective assessment and objective indicator. Institutions are not given an individual ranking position but assigned to a ranking group of top, middle and end group – which gives the appearance of a traffic light presentation (see Fig. 5.7). A comparable approach has been adopted in the food industry where 'traffic light' graphics are used in food packaging to inform the consumers on food quality.

The ranking, therefore, never tells the user who is the best but maybe who is performing better than average on indicators considered relevant to the user.

The CHE Research Ranking (CHE-Forschungs Ranking) currently covers 16 subjects from natural sciences, humanities and social sciences. It does not define ranking positions but determines the top groups for individual indicators. It is determined based on the following factors:

- Level of third-party funding spent on individual subjects
- Number of doctorates
- Publications and citations
- Patent registrations or inventions

Interviews with professors provide additional information that is given on the reputation of universities with respect to the subjects analysed. However, this information is not used to determine the top groups.

5.5 Demystification and Confidence

The challenge set for this chapter was to explore the methodologies used by the compilers of university, college, and research rankings and to test out to what extent we should have confidence in them.

The view taken early in the chapter was that a negative critique of the methodology does not imply an intention to mislead the reader of such rankings. It was suggested that the ranking methodology is complex and occasionally, compilers are reluctant to de-mystify steps used – for commercial sensitive reasons – and thus, we the users of rankings have to rely on the validity, reliability and consistency of the output from the methodology adopted and applied by the compilers and publishers. Leach (2004), from the perspective of the compilers, comments on the limitations of the university rankings.

> University table or more specifically the rankings we employ, generate a fair amount of anger in the academic community. Institutions are often annoyed at the methodology and the data we choose, and at the sheer gall of marking them against each other in the first place. But we believe that, on balance, tables like these are important. (Leach 2004)

It is clear from the quotation that Leach feels that '… on balance…', there is more to be gained from the tables than lost and as the impact of debt increases, it is important that students '…know what they are getting for their cash'. It is my proposition that the tables really do not provide the answer that they may be searching for.

The *black box* has been opened, the compilers have made available insights into the processes they perform to create the index, yet most of us are unwilling – not unable – to engage in a critical discourse with the compilers to challenge them to provide a justification for each step in the process and to provide a philosophically sound rational justification that allows them to use a single metric to define a university.

The final most critical question remains 'How can a university be reduced to a single metric which is meaningful?'

It remains the single most disconcerting aspect of the whole process of creating a ranking, one that defies logic and one that is so patently wrong. A university is a complex, dynamic organisation constantly changing, year on year with respect to the faculty providing the teaching, to the form and nature of the curriculum offered, to the resources provided. To capture all that complexity in a single measure makes little sense.

Add to this the fundamental methodological criticism described in the paragraphs above where at each stage in the process profound criticism have been advanced at the limited theoretical framework informing assumptions adopted by compilers.

- From the selection of specific events over other specific events
- Their conversion into numerical values
- The adaptation of these numerical values on to scales
- The aggregation of these scaled indices to create a single measure
- A theoretical belief that the measure is capable of defining the quality of a university, a teaching subject, a department, or a research group

Universities are complicit in the process and fail to critically stand out for a more robust and honest attempt at providing information to prospective students rather than play '*our ranking is better than your ranking*'.

A critic of *US News* and *World Report*, Thompson (2000) claims that there is sound evidence that universities and colleges alter their policies for the sake of the rankings – the Heisenberg effect, thus changing the very thing being measured – and giving rise to the danger of mission drift, valuing aspects of university and college life that are exposed to the measurements and thus devaluing those aspects less open to an objective measurement.

Thompson maintains that rankings are:

> … opaque enough that no one outside the magazine can figure out exactly how they work, yet clear enough to imply legitimacy.

A view that accurately summarises the position in 2010.

References

Altbach, P. (2006). The dilemmas of ranking. *International Higher Education*. Retrieved 25 June 2009, from http://www.bc.edu/bc_org/avp/soe/cihe/newsletter/Number42/p2_Altbach.htm

Baty, P. (2009). The world in motion. *Times Higher Education*. December 17, 2009.

Brown, R. (2006). League tables – do we have to live with them? *Perspectives: Policy and practice in higher education, 10*(2), 33–38.

Casper, G. (1997). Private letter from Gerhard Casper, president of Stanford University, to James Fallows, editor of U.S. News & World Report, Personal communication received by J. Fallows, on September 23, 1996.

Cave, M., Hanney, S., Henkel, M., & Kogan, M. (1997). *The use of performance indicators in higher education: the challenge of the quality movement*. London: Jessica Kingsley.

Centre for Higher Education Development. (2009). Retrieved 25 July 2009, from http://www.che-ranking.de/cms/?getObject=613&getLang=en

Clarke, M. (2002). Quantifying quality: What can the U.S. News and World Report ranking tell us about the quality of higher education? *Education Policy Analysis Archives*. Retrieved 20 June 2009, from http://epaa.asu.edu/epaa/v10n16

Clarke, M. (2004). Weighing things up: A closer look at US News & World Report ranking formulas. *College and University Journal, 79*(3), 3–9.

Dwyer, M. (2008). Our 18th annual rankings. Retrieved 4 Dec 2009, from http://oncampus.macleans.ca/education/2008/12/19/our-18th-annual-rankings/

Ehrenberg, R. G. (2003). Reaching for the brass ring: The U.S. News & World Report rankings and competition. *The Review of Higher Education, 26*(2), 145–162.

Graham, A., & Thompson, N. (2001). Broken ranks. Article in *Washington Monthly*.

Harvey, L. (2008). Rankings of higher education Institutions: A critical review. *Quality in Higher Education, 14*(3), 187–208.

HESA. (2009). Student definitions 2007/08. HESA Student Record Retrieved 20 June 2009, from http://www.hesa.ac.uk/index.php/component/option,com_datatables/task,show_file/defs,1/Itemid,121/catdex,3/disp,/dld,institution0708.xls/yrStr,2007+to+2008/dfile,studefs0708.htm/area,institution/mx,0/

Hinton, P. (2004). *Statistics explained*. Hove, East Sussex, UK: Routledge.

Leach, J. (2004). How not to use the tables. Article in *The Guardian*.

Lederman, D. (2009a). Manipulating,‘ Er, Influencing ’*U.S. News*. *Inside Higher Ed*. Retrieved 22 June 2009, from http://www.insidehighered.com/news/2009/06/03/rankings

Lederman, D. (2009b). More rankings riggings. *Inside Higher Ed*. Retrieved 22 June 2009, from http://www.insidehighered.com/news/2009/06/08/usc

Longden, B. (2008). Performance Indicators. In G. McCulloch & D. Crook (Eds.), *The Routledge International Encyclopaedia of Education* (pp. 760).

Longden, B., & Yorke, M. (2009). Institutional rankings, marketing, and the needs of intending students. In B. M. Kehm & B. Stensaker (Eds.), *University rankings, diversity, and the new landscape of higher education*. Rotterdam: Sense Publications.

MacLeod, D., & Hiely-Rayner, M. (2007). University tables 2008: Methodology. *Guardian*. Retrieved 1 June 2009, from http://education.guardian.co.uk/universityguide2008/story/0,,2067150,00.html

MacLeod, D., & Hiely-Rayner, M. (2008). University tables 2009: Methodology. *Guardian*. Retrieved 1 June 2009, from http://education.guardian.co.uk/universityguide2009/story/0,,2067150,00.html

Macleod, D., & Hiely-Rayner, M. (2009a). University guide 2009: The calculations behind the tables. *Guardian*. Retrieved 2 June 2009, from http://education.guardian.co.uk/university2009/story/0,,2276943,00.html

MacLeod, D., & Hiely-Rayner, M. (2009b). University guide 2010: University league tables. *Guardian*. Retrieved 20 June 2009, from http://www.guardian.co.uk/education/table/2009/may/12/university-league-table

Miles, J., & Shevlin, M. (2006). *Applying regression & correlation: A guide for students and researchers*. London: Sage Publications.

Morris, H. (2005). A rank influence. Article in *The Guardian*.

Morse, R. (2009). Undergraduate ranking criteria and weights. Retrieved 29 June 2009, from http://www.usnews.com/articles/education/best-colleges/2008/08/21/undergraduate-ranking-criteria-and-weights.html?PageNr=3

Morse, R., & Flanigan, S. (2006). America’s best colleges 2007. Retrieved 8 May 2007, from http://www.usnews.com/usnews/edu/college/rankings/about/weight_brief.php

Morse, R., & Flanigan, S. (2009). How we calculate the rankings. Retrieved 22 June 2009, from http://www.usnews.com/articles/education/best-colleges/2008/08/21/undergraduate-ranking-criteria-and-weights.html?loomia_ow=t0:s0:a41:g2:r5:c0.212928:b20174593:z0&s_cid=loomia:about-the-rankingsmethodology

Newman, M. (2008). Students urged to inflate national survey marks to improve job options. *The Times Higher*. Retrieved 25 June 2009, from http://www.timeshighereducation.co.uk/story.asp?sectioncode=26&storycode=401883

O'Leary, J., Hindmarsh, A., & Kingston, B. (2006). *The good university guide 2007*. London: The Times.
QS. (2009). QS.com Asian university rankings 2009. From http://www.qsnetwork.com/
Richardson, J. (2008). *Counting what is measured or measuring what counts? – League tables and their impact on higher education institutions in England*. Bristol: Higher Education Funding Council for England: Appendix B.
Richmond, W. K. (1968). *Readings in education*. London: Methuen & Co Ltd.
Ryan, A. (2009). The trouble with tables. Article in *The Times Higher.*
Shapiro, B. (1994). Personal communication from Bernard Shapiro to Ann Dowsett, Personal communication received by A. Dowsett, on 20th September 1994.
Shea, C. (2009). Shocked, shocked: More evidence of rankings shenanigans. *Boston Globe*. Retrieved 4 Dec 2009, from http://www.boston.com/bostonglobe/ideas/brainiac/2009/06/shocked_shocked.html
Stella, A., & Woodhouse, D. (2006). Ranking of higher education institutions. J. Baird: AUQA Occasional Publications Series no. 6. Melbourne: AUQA from http://www.auqa.edu.au/files/publications/ranking_of_higher_education_institutions_final.pdf
Thompson, N. (2000). Playing with numbers: How U.S. News mismeasures higher education and what we can do about it *Washington Monthly*.
Times online. (2009). Good university guide 2010: How the tables work. *Times Online*. Retrieved 20 June 2009, from http://extras.timesonline.co.uk/tol_gug/gooduniversityguide.php
Usher, A., & Savino, M. (2007). A global survey of rankings and league tables. In Institute for Higher Education Policy (Ed.), *College and university ranking system*. Washington, DC: Institute for Higher Education Policy.
Watson, D. (2008). Universities behaving badly. *Higher Education Review, 40*(3), 3–14.
Webster, D. S. (1986). *Academic quality rankings of American colleges and universities*. Springfield, IL: Charles C. Thomas.
Yorke, M., & Longden, B. (2005). *Significant figures: Performance indicators and 'league tables'*. London: Standing Conference of Principals.