The Structure of Surveys and the Peril of Panels

Adrian Simpson
School of Education
Durham University
Leazes Road
Durham, DH1 1TA, UK.
adrian.simpson@durham.ac.uk

September 26, 2016

Abstract

University league tables give the image that there is a single dimension along which institutions can be placed. Most derive rankings from an aggregate score of multiple items, which often include opinion responses which has the potential to introduce sampling bias. This paper explores what happens when the two issues of dimensionality and sampling bias interact in league tables. It uses the *Times Higher Education* Student Experience Survey — which uses a panel design — to explore these issues. It notes that they combine to produce a distorted image of the relative quality of the student experience in different institutions. We conclude that ignoring dimensionality and the systematically unrepresentative nature of the sample could lead policy makers to draw inappropriate conclusions.

1 Introduction

University league tables have become a familiar part of higher education. While their introduction may date back to the 1920s (Salmi & Saroyan, 2007), the increasing number of them and their apparently increased importance coincides with an increased commodification of higher education and particularly, as the market for high (and less regulated) fee paying international students has grown (Harvey, 2008).

Much existing research has focussed on international comparisons (such as the *Times Higher Education World University Rankings* or the *Academic Ranking of World Universities*). That research notes serious methodological issues including:

- Incompatability of institutions
- Arbitrary aggregation weightings
- Methodological tinkering
- The role of opinions
- The lack of error bars
- Consequential validity

Many league tables seek to compare institutions across broad categories. Some compare across countries, while others compare institutions within disciplines or within a single country. Concern has been raised that there is no single conception of quality which makes ordinal comparison possible between institutions with very different missions, embedded in different cultures and with different patterns of disciplines. Proulx (2007) argues that rating

systems should only 'compare comparables'; that does not only mean restricting inclusion to institutions with similar missions, but also ensuring comparability at the level of disciplines. Proulx goes on to suggest that overall rankings should be avoided altogether.

In addressing how overall rankings are obtained from the aggregation of items, Usher and Savino (2007) list 17 different league tables which use between 1 and 71 different indicators. They separate indicators according to source (survey, third party data and university sources) and category (input and output) and note that the weightings between the indicators vary widely between them. van der Wende and Don (2009) argue that the weightings are arbitrary and lack theoretical foundation and Usher and Savino (2007) suggest that the poor consistency between weightings and rankings means that league tables are not measuring what is intended (institutional quality) but that the indicators are epiphenomena of other unmeasured aspects of universities (such as age or faculty size). There is also concern about the differences in choices of scale and variance of measures. While Williams and de Rassenfosse (2016) argue that if things "are similar in one area but substantially different in another then this should be allowed to affect the rankings and not be standardised away" (p.7), differences in scale do matter: constructing a ranking by averaging research income (in, say, dollars) with percentage of international staff (out of 100) without standardising in some way would result in rankings being almost entirely representative of research income.

Salmi and Saroyan (2007) note the problems caused by compilers modifying the criteria, weightings and other methodological factors year-on-year. They exemplify this by referring to a change in methods used for the 2005 THE World University Rankings: "Malaysia's top two universities ... [slipped] ... by almost 100 places compared to the previous year. In response, the leader of the opposition called for a Royal Commission of Inquiry, notwithstanding the fact that the dramatic decline was partly due to a change in the ranking methodology" (p.10). Such methodological adjustment always has a historico-political dimension: Yorke (1997) suggests that when one league table found that a new survey item ('cost of living') led to very large changes in 'expected' league table positions, they decreased the weighting to try to maintain some element of the status quo.

Among these various indicators, many league tables rely on measures of reputation which come from responses to opinion polls conducted with, for example, graduate recruiters or academics at other institutions. Concerns are raised that many taking part in the opinion polls may have insufficient knowledge to make an accurate assessment and that the polls may be reflective of a longer term, even historical, view of the standing of the institution, rather than its current performance (Dill & Soo, 2005). There is even evidence that respondents can give high ratings for departments that don't exist (Brooks, 2005). By averaging and aggregating responses of what are, inevitably, subjective opinions and reporting them, with apparently high precision, to many significant figures, pollsters can give the image of objectivity.

Particularly when responses rely on sampling techniques, one would expect to have some level of sampling error, even if there are no other sources of randomness. The rank ordering of those responses, then, cannot be definite: one institution may have a slightly higher score than another as a result of the samples chosen, not because it has some objectively higher score. Goldstein and Spiegelhalter (1996) argue that ranking without accounting for uncertainty can lead to drawing incorrect conclusions (including, for example, rationalising large changes in positions which may be no more than random fluctuation or regression to the mean). van der Wende and Don (2009) suggest that for this, amongst other reasons, rank orderings should be avoided.

Messick (1995) argues that one concern we should have about an instrument is in the consequences it has for policy: for example, as a result of an unexpectedly low league table ranking, an institution might choose to radically alter aspects of how it is organised. This may be a valid decision, but if the low ranking came from no more than a compiler's methodological change or a quirk of sampling error, the policy decision might not be valid. Even where rank ordering may have both conceptual validity and statistical significance, the

consequences of basing strategic decision making on ranking positions can lead to reduced diversity of mission and the homogenisation of institutions (Shin & Toutkoushian, 2011). It is clear that institutions do take league tables seriously: Hazelkorn (2008) suggests that a large proportion of senior managers use rankings to influence their strategic decisions, even though the evidence is mixed about the extent to which students use them (Clarke, 2007; Gibbons, Neumayer, & Perkins, 2015). Moreover, even when surveys take great care to report nuanced and hedged findings and discourage the simplicity of a hierarchy, the message from the media is often simply focussed on ranking position (Blasi, Romagnosi, & Bonaccorsi, 2016).

This paper is not intended to revisit these concerns, but to draw out two interacting issues which have received less attention in discussions of league tables: dimensionality and systematic sampling bias. It explores some issues related to data when it is presented as representing a single dimension of 'quality', whether unidimensionality is justified, what happens when (intentionally or not) sample selection co-varies with an important survey measure and, finally, what happens when systematic sample bias and multi-dimensionality interact. It uses one published league table (the *Times Higher Education Student Experience Survey*) as a case study, not because other league tables are free of these issues, but because they appear in stark relief in this particular survey.

2 A Case Study: the student experience survey

The *Times Higher Education* (a British weekly magazine focussed on the university sector) has published an annual student experience survey of higher education institutions in the UK for over a decade. It aims to be "relevant not only to those providing the student experience but to those advising future cohorts of students as well ... [and] ... to provide teachers with the information they need to help their students as they consider their options for post-secondary education" (Gill, 2014, p.3). The survey is conducted by a market research company, trading as 'YouthSight'.

The 2016 survey was based on the responses from 15000 full time undergraduates who are part of the YouthSight's student panel (Briggs, 2016). These are students who have agreed to take part in the company's surveys, receiving a £25 voucher for undertaking between 7 and 15 surveys, each taking around 10 minutes (YouthSight, 2015).

The survey is based on 22 questions. The first 21 involve rating various aspects of the student experience on a seven point scale. These include attributes like 'good extra-curricular activities/societies', 'good industry connections' and 'tuition in small groups'. The final question is not included in subsequent aggregations and asks respondents for their level of agreement with the statement 'I would recommend my university to a friend'. The first 21 questions are weighted (by multiplying the average response for the institution by 2, 1.5 or 1), summed and then scaled out of 100 as the maximum possible score. The aggregate score is then used to create a rank ordering of institutions. The labels in figure 1 list the attributes and their weightings. Clear details about the panel and the methods are given across publications from YouthSight and THE (Briggs, 2016; YouthSight, 2015).

Unlike many university league tables, this is based entirely on survey data and, unlike rankings which include opinions about reputation from those working at other institutions (like the *Academic Ranking of World Universities*), the survey is conducted with participants who are studying at the institutions. It may be less likely, then, to suffer from the issue of rating historical reputation rather than current performance (Dill & Soo, 2005). However, the vast majority of students only ever have experience of a single higher education institution, so it is unclear how well they can make a judgement on, say, the quality of their students' union, in the absence of being able to compare with other students' unions.

As with most university ranking systems, the survey is open to many of the criticisms listed above; but this paper focusses primarily on issues of dimensionality and the impact

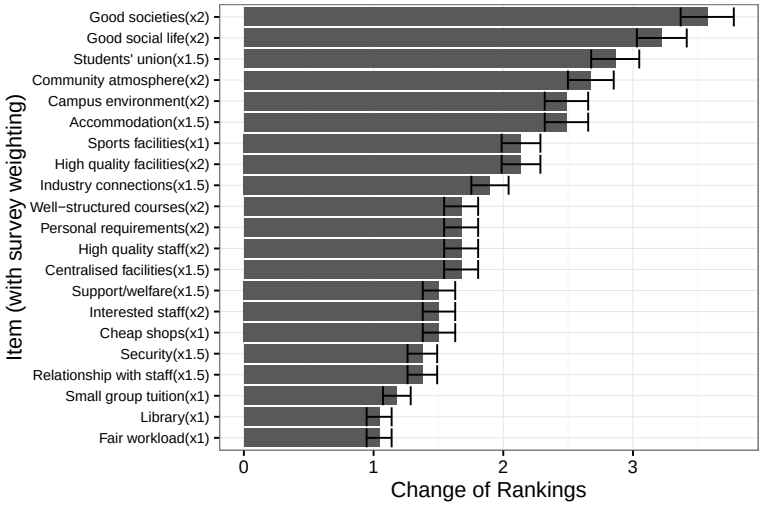


Figure 1: Mean impact of 1σ increase in a given item

of the panel sampling on multidimensionality. We begin by looking at the uni-dimensional structure of the survey.

3 Unidimensional structure: Weighting Analysis

It has been repeatedly noted that the weightings used to aggregate the factors to construct a single rating in league tables can appear arbitrary and lack theoretical foundation (Daraio, Bonaccorsi, & Simar, 2015; Dill & Soo, 2005; Ramsden & Callendar, 2014).

In the student experience survey, the various items are weighted presumably to reflect what the designers believe to be more and less important aspects of the student experience. In earlier versions of the survey, there is evidence that they have altered weightings: for example, in 2009, the weighting for 'good security' and 'good accommodation' were increased on the basis of a higher correlation with the question about whether the participant would recommend the institution to a friend (Attwood, 2009).

Yorke (1997), however, noted that the relationship between weighting and rank ordering may not be that simple: all other things being equal, items with a higher standard deviation will have more impact on ranking than items with a lower standard deviation (though weighting and standard deviation will interact). Figure 1 shows the mean impact of a one standard deviation increase in each variable (averaged across all institutions) holding all other variables the same.

This suggests that the impact of an item does not fit well with the weightings: a 1σ increase in the response of 'good students' union' or 'good accommodation' (without changing any other responses) has a bigger average impact on rank position than, say, a 1σ increase in 'interested staff' (leaving other responses unchanged), even though the last item was intended to have the higher weighting. This is a consequence of a failure to standardise the individual item measures before aggregating them (Longden, 2011; Soh, 2013).

Moreover, on aggregating, the choice of weighting can directly affect rankings. In relation to international league tables like the THE *World University Rankings* or the *Academic Ranking of World Universities*, Saisana and d'Hombres (2008) undertook an uncertainty and sensitivity analysis. That is, they looked at how the rankings altered depending on the extend to which different methodological assumptions are made. They found that rank

Table 1: Alternative weighting

Weights		
1	1.5	2
High quality staff	Support/welfare	Students' union
Interested staff	Accommodation	Centralised facilities
Well-structured courses		Industry connections
Good social life		Cheap shops
Community atmosphere		Library
Good societies		Sports facilities
Campus environment		
High quality facilities		
Personal requirements		
Relationship with staff		
Security		
Small group tuition		
Fair workload		

orderings varied widely depending on weightings and inclusion or exclusion of indicators. They recommend that such an analysis should be undertaken for all league tables constructed in this way.

A similar analysis was therefore developed for the Student Experience Survey data. The aim was to assess the extent to which ranking varied according to the weighting chosen for items. While it is possible to do this with entirely random weights by sampling weighting vectors from the 20 dimensional simplex, a more approachable method (which gives very similar results) is to vary the weights between the three values used in the analysis of the survey. To this end 10000 (of the over one billion) different ways of allocating the weights 1, 1.5 and 2 to the 21 different items were uniformly sampled and applied to the reported item scores in the survey. Figure 2 shows boxplots of the results (with the box representing the middle 50% of resulting ranking positions of the given institution and the whiskers extending 1.5 times the interquartile range) along with a cross representing the ranking position in the published survey.

The figure shows that the ranking, particularly away from the extremes, is highly dependent on the choice of weighting. For example, if the weightings were as in table 1, then Bristol (ranked 48th in the published survey, 10 places above Central Lancashire) would drop to 71st (19 places below Central Lancashire with the alternative weightings). Note that figure 2 only shows how sensitive ranks are to different choices of weightings, it does not show how wide the confidence intervals of the rankings should be if we accounted for sampling error (which is likely to be so wide as to make even crude separation of institutions impossible and thus eliminate the justification for publishing the survey at all; see Goldstein & Spiegelhalter, 1996).

It is also clear from comparing the median of the simulated data (indicated by the vertical bar in each box) with the published rank that some universities are advantaged and some disadvantaged by the chosen weighting (that is Bristol's published ranking is 8 places higher than its median in the simulation and Central Lancashire is 3 places lower). We note that in general, the published weighting tends to disadvantage institutions with larger proportions of state school entry: there is a strong negative correlation between the difference between the published and median simulated ranking with the proportion of state school pupils ($r^2 = 0.26, p < 0.0001$). In contrast, with the weighting in table 1 there is only a very modest negative correlation with the proportion of state school pupils ($r^2 = 0.07, p = 0.04$).

van der Wende and Don (2009) note an alternative to compilers' ungrounded choice of weightings. In the German Centre for Higher Education interactive online system, students can choose their own criteria and, rather than a single number or rank position, the website displays a profile of universities against those criteria. It is interesting to speculate how the THE student experience survey might work under a similar user-led weighting system. A

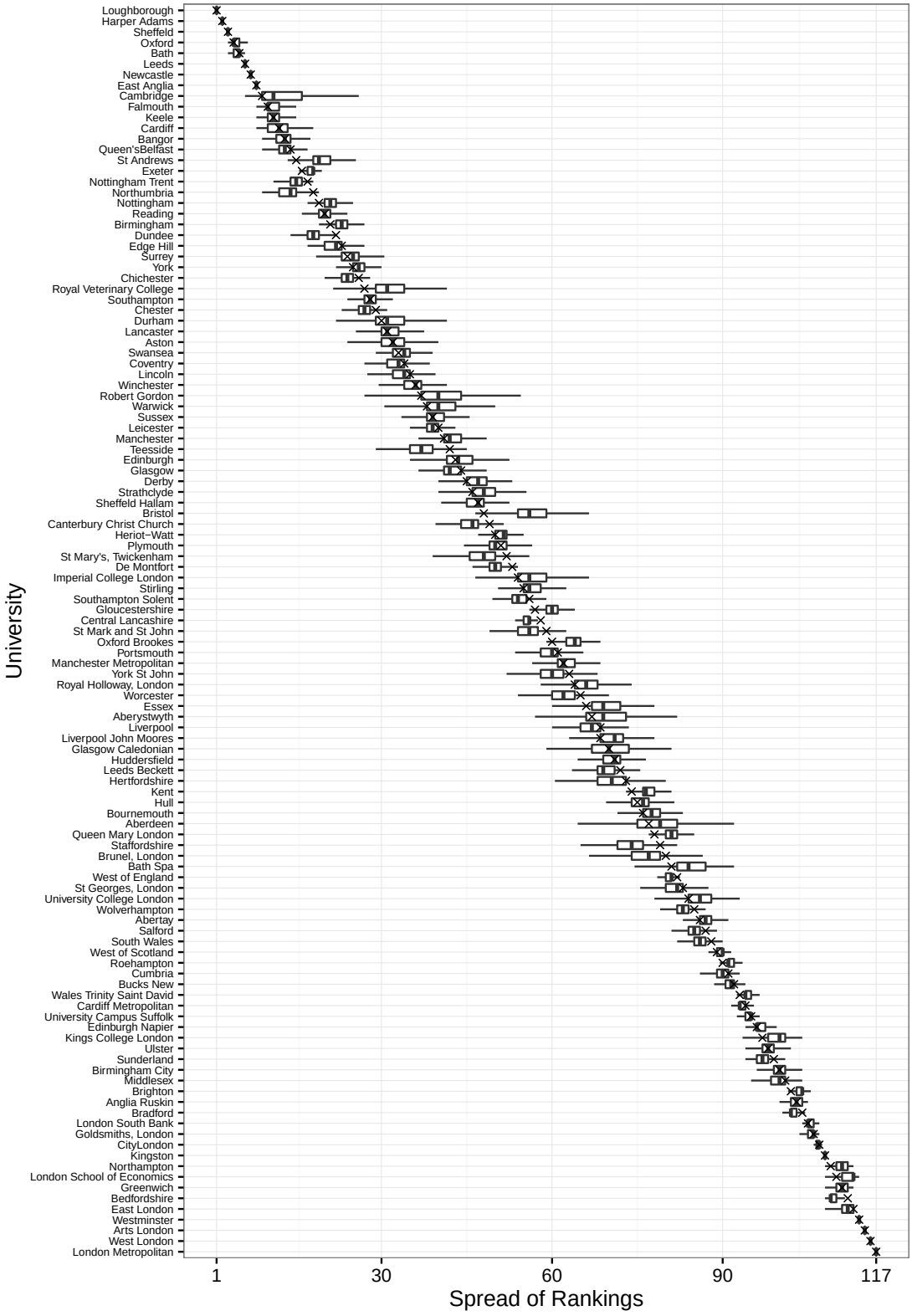


Figure 2: Spread of ranking positions depending on weighting.

student who shares the views of the survey compilers and weighted the items as in the published table (indicated in figure 2) would see Bristol ranked well above Central Lancashire. Another student who weighted facilities more importantly than staffing and social activities might weight in a similar manner to table 1 and would see Central Lancashire ranked well above Bristol. Both might be entirely legitimate and might reflect the different interests and needs of the students. Of course, it is easier for a commercial publisher to sell their product if they can declare a single 'winner'.

4 Multi-dimensional structure

However, using any weighting of items to aggregate averaged responses to a single rating (from which a rank ordering, and a 'winner' can be derived) presupposes that institutions can be ordered on a single dimension representing some platonic sense of 'quality'.

In order to see if it is justified to consider the data as all pointing to a single underlying latent variable, it is possible to explore its dimensionality. To identify the component structure of the data, we undertook a principal component analysis, with oblique axis rotation (oblimin) on the 21 item scores. Kaiser-Mayer-Olkin's measure of sampling adequacy suggested the analysis should yield reliable and distinct components (with the overall measure = 0.91). Bartlett's test of sphericity was highly significant ($\chi^2(210) = 2501, p < 0.0001$) indicating that the inter-item correlations were large enough for the analysis.

Rather than the data being well described by a single component, parallel analysis indicated that 2 components would be most appropriate. Table 2 shows the factor loadings (greater than 0.55) after rotation.

Table 2: Principal component analysis

Variable	PC1	PC2	h2	u2	com
Good societies	**0.93**		0.82	0.18	1.00
Centralised facilities	**0.92**		0.71	0.29	1.11
Good social life	**0.88**		0.73	0.27	1.01
Accommodation	**0.80**		0.73	0.27	1.04
Sports facilities	**0.80**		0.59	0.41	1.01
Community atmosphere	**0.78**		0.79	0.21	1.14
Campus environment	**0.77**		0.80	0.20	1.18
Students' union	**0.73**		0.43	0.57	1.24
Library	**0.71**		0.46	0.54	1.03
High quality facilities	**0.71**		0.69	0.31	1.20
Support/welfare	**0.66**		0.82	0.18	1.60
Security	**0.62**		0.58	0.42	1.31
Cheap shops	**0.59**		0.38	0.62	1.02
Industry connections			0.25	0.75	1.89
Relationship with staff		**0.97**	0.78	0.22	1.16
Interested staff		**0.87**	0.90	0.10	1.06
Small group tuition		**0.82**	0.77	0.23	1.04
High quality staff		**0.73**	0.83	0.17	1.35
Personal requirements		**0.56**	0.80	0.20	1.96
Well-structured courses		**0.56**	0.81	0.19	1.97
Fair workload			0.11	0.89	1.55
SS loadings	9.04	4.75			
Proportion Var	0.43	0.23			
Cumulative Var	0.43	0.66			
Cum. factor Var	0.66	1			
PC1	1.00	0.45			
PC2	0.45	1.00			

In addition to their statistical derivation, these components make sound theoretical sense:

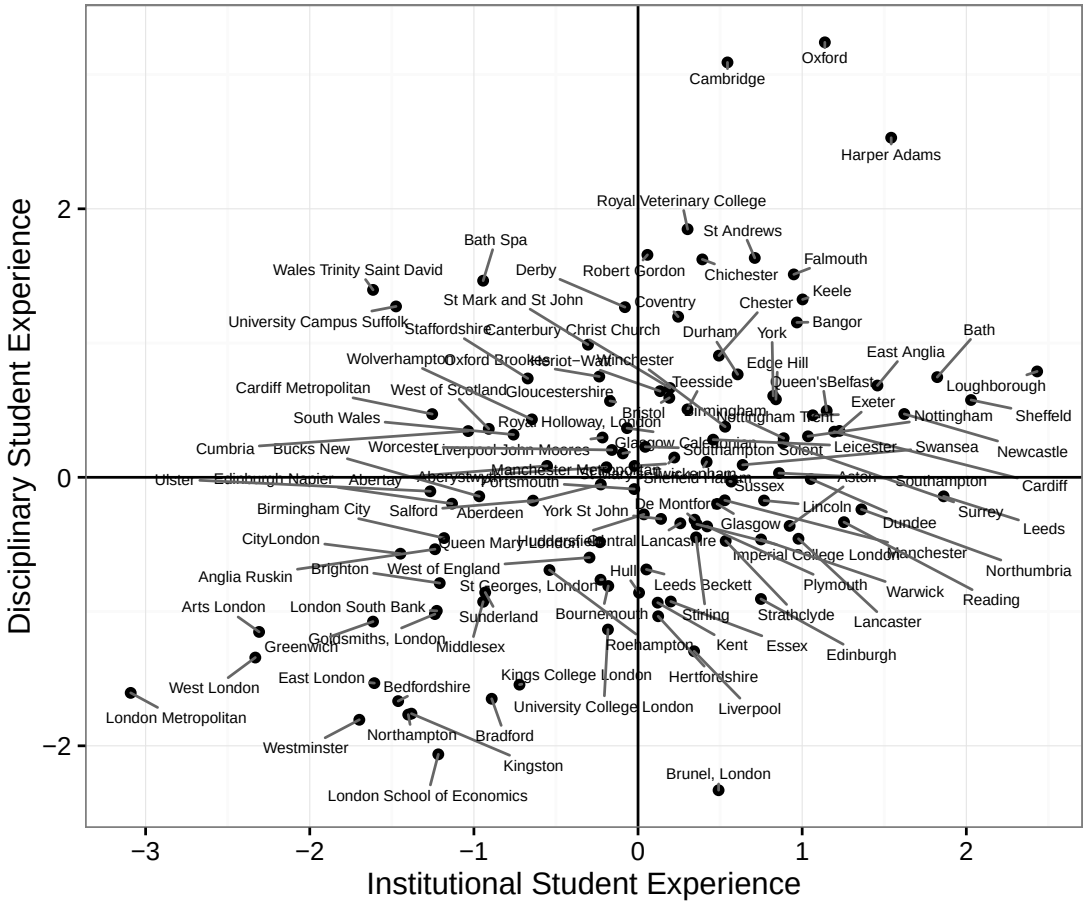


Figure 3: Scatterplot of two principal component solution of institution.

The first component draws together items which would generally be organised at the institutional level such as the accommodation, environment and sports facilities. The second draws together items which would be expected to be organised by discipline including teaching staff, course structure and tuition group size. Thus, we could argue that the variance in the responses to the survey are best described by two components: one measures the *institutional student experience* and one measures the *disciplinary student experience*.

Figure 3 shows the institutions included in the original survey with their component scores achieved in against these two dimensions. It suggests that institutions which may sit at relatively modest positions in the published league table, such as Robert Gordon University (published rank 38), may have strong perceived disciplinary student experience which is masked by somewhat more modest institutional student experience. Alternatively, those with strong league table positions may have achieved this from being perceived to be outstanding in one dimension while being less well regarded by respondents in the other.

Note the aggregating of multiple dimensions to one can also cause instances of *Simpson's Paradox*: when we aggregate the scores of all 21 items to just one rank ordering we can find one university ahead of another, but when we aggregate to two dimensions separately, the second can outscore the first on both dimensions. For example, in figure 2 Aberystwyth outranks Glasgow Caledonian on the published survey (and, indeed, would do so even if the

weighting was equal across all items), yet figure 3 shows that Aberystwyth has lower scores for both institutional student experience and disciplinary student experience. For a fuller explanation of Simpson's Paradox and how commonly it occurs when we aggregate data to obtain rankings, see Haunsperger (2003)

This separation of the student experience survey into two components draws in to sharp relief the question of the sample. James MacGregor, Director of Higher Education for YouthSight, acknowledges concerns with sample size, stating "Although the number of respondents comprises less than 1 per cent of the UK full-time undergraduate population, this sampling fraction is high in comparison with a typical political opinion poll or large-scale government survey. More importantly, the overall sample size is large enough to generate only a small sampling error" (Briggs, 2016, p. 15). While 1% of a very large population may give a reasonably good estimate of some characteristic of the population, political opinion polls and surveys are often looking for differences between a small number of options and will note statistical 'dead heats' when differences are within a specified margin of error. In this case, we have 117 different institutions and no measure of margin of error provided. In addition to these concerns, we need to consider two issues: the nature of the population of which the sample is said to be representative of a given characteristic and the nature of the sampling error.

One immediate consequence of noting the survey's conflation of two components is that it highlights the issue of sampling across those two components. Consider an institution of 10000 students where 100 of them participate in the survey. That institution probably has one set of sports facilities shared by those 10000 students (and therefore by the sample of 100). The response from the sample will be expected to be *uni-modal*: that is, distributed around some single value which one can then take to be a point estimate of the collective judgement of the sample on that shared set of facilities.

However, the sample of 100 students may contain 5 historians, 20 engineers, 10 mathematicians, etc. Their views of questions about lecturing staff will be *multi-modal*; it will not be distributed around a single value, but each disciplinary subgroup's responses will be distributed about a value for their set of teaching staff which may not overlap with other disciplinary subgroups in the sample. Even if one thought it made sense to talk about the quality of teaching staff at an institution (as distinct from the quality of the staff in each department) any such institutional component will be poorly represented by the sample, given how few survey participants one would expect to come from each department and how widely those small sample sizes would vary within and between institutions. In particular, this means that, ceteris paribus, one would expect the variance of the responses on the disciplinary student experience questions to be much wider than the variance of the responses on the institutional student experience.

Contrast this with the UK's National Student Survey (NSS) approach which requires a 50% response rate from each discipline (with a minimum response of 23) to reach the publication threshold. This ensures a suitable level of precision at the disciplinary level, albeit that there is evidence of quite different response patterns, depending on discipline (Yorke, Orr, & Blair, 2014). Arguably, the NSS also contains questions which are institutional level (such as those about resources) alongside those which are disciplinary (such as those about teaching). However, the main mechanism by which the NSS ranks institutions is not a simple aggregate of these questions, but a separate overall satisfaction question (although the *Times Good University Guide* does include a simple average of the NSS items without considering dimensionality, Cheng & Marsh, 2010).

5 Sampling bias

While the sample size may vary in our view of its adequacy across the two dimensions of student experience, it also appears that there is systematic sampling bias.

The survey compilers give a range of outline demographics about their student panel,

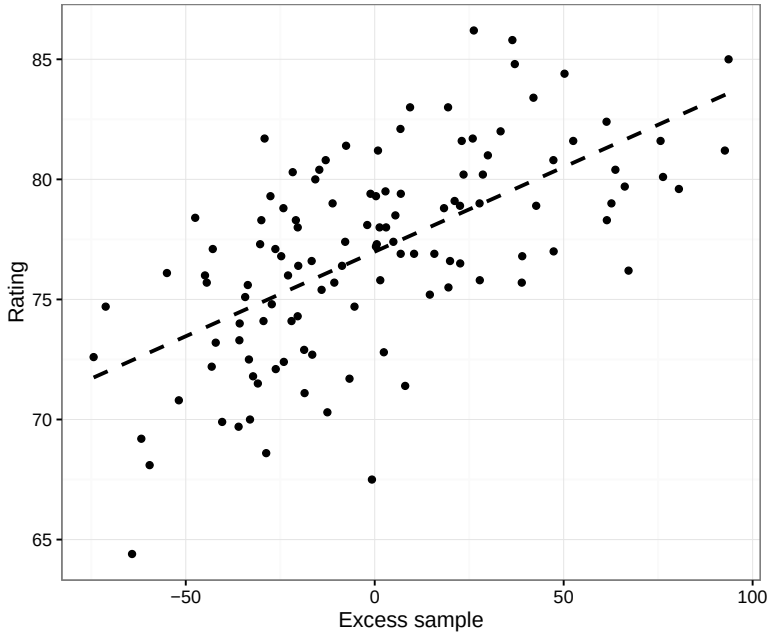
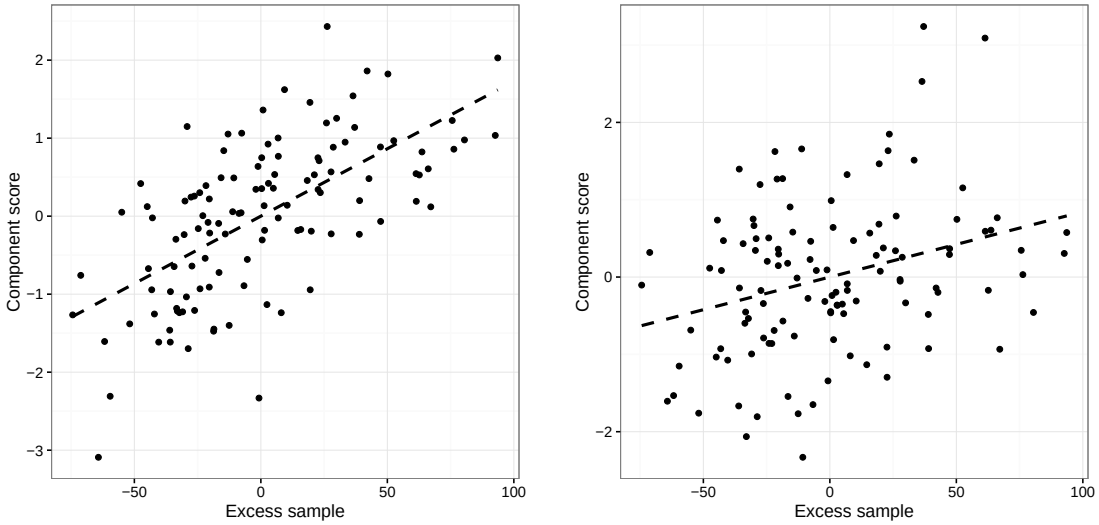


Figure 4: Relationship between the excess sample size and rating.

which consists of over 72000 students at UK higher education institutions (YouthSight, 2015). This represents between 3 and 5% of the UK student population (depending on whether it is intended to represent only full time undergraduates, or include postgraduate and part time students as well). It seems to over-represent female students: They form 62% of the panel, but the Higher Education Statistics Agency indicates that between 54 and 56% of the population are women (again, depending on what is being counted). It also over-represents students at Russell Group universities (a group of 24 UK institutions which brand themselves as "research-intensive, world-class"), from state school backgrounds etc. Given that, as mentioned above, the panel is an opt-in process, for which there is only a small financial incentive to take part, it is unsurprising that the pattern of representation on the panel does not match that of the student population at large. Such sampling bias is likely to be present elsewhere: Porter and Umbach (2006) suggest that the US National Survey of Student Engagement, oversamples from high ability students, women and white students.

However, one more surprising sampling bias shows up when the sample size in the student experience survey is compared to the aggregated rating achieved. There is a large correlation between the two ($r^2 = 0.25, p < 0.0001$). It might be argued that this may be the result of an obvious potential confound: that the sample size varies with the institution size (which it does) and that there is something about being a bigger institution which leads to more positive responses about student experience. However, this latter is not substantially the case: there is a very small, not statistically significant relationship ($r^2 = 0.001, p = 0.29$). Moreover, when one adjusts for the relationship between institution size and sample size – that is, looks at the size of the sample above or below that expected for the institution size – the relationship with rating is even stronger ($r^2 = 0.38, p < 0.0001$). Figure 4 plots the rating against the number of students in the sample above or below that we would expect for the size of the institution and shows the strength of this residual relationship.

In a unidimensional setting, this sampling bias, while worrying, might not necessarily



(a) Institutional Student Experience

(b) Disciplinary Student Experience

Figure 5: Relationship between sample size and component scores

lead to serious flaws in the interpretation of the findings or upset the rank ordering (we will see below that the bias tends to increase all the results on a unidimensional scale of ratings beyond those we would expect from a random sample and squeeze them together, but not necessarily to reorder them). However, in a multi-dimensional scale, where a rating is calculated by aggregating items which may not form a single coherent scale, the relationship between the sampling bias and the different scales may be unequal and this may cause more serious problems.

6 Sampling bias and structure

This unbalanced relationship is present in the published survey data. Around 40% of the variance in the sampling excess (after accounting for institution size) is shared with variance in the institutional student experience scale ($r^2 = 0.40, p < 0.0001$, see figure 5a), but less than 10% is shared with the disciplinary student experience ($r^2 = 0.09, p = 0.0007$, see figure 5b). Indeed, even this 10% of shared variance may be due to the remaining correlation between the disciplinary and institutional student experience components.

That is, positive views of the institutional student experience (such as 'extra-curricular activities' and 'good social life') vary closely with participation in the survey; while positive views of disciplinary student experience (such as 'relationship with staff') seem relatively independent of participation in the survey.

Suppose that students are more likely to agree to be a part of the panel and to take part in the survey if they are more enthusiastic, have more time on their hands (and so, perhaps be engaged with extra-curricular activities more than average) or otherwise be skewed to respond more positively to aspects of their institutional student experience than a genuinely random sample of their peers. This has the result of tending to inflate the sample estimate of the institutional student experience, while leaving the disciplinary student experience estimate relatively unchanged.

To demonstrate this, a simulation can be constructed, illustrating two fictional institutions (University A and University B) each of which has, say, 5000 students.

Each grey dot in figure 6 represents the response one student would give for the two aspects of their experience (an institutional aspect and a disciplinary one). So, for each institution there is a cloud of 5000 grey dots representing those responses. The large black circle represents the centre (centroid) of that cloud. This, in effect, represents the rating the institution would receive if every student participated in the survey. Note that, when we look at the whole population, the mean institutional student experience is higher for University A than University B while the mean disciplinary student experience is very slightly lower for A than B. If, as the real survey does, we simply aggregate these two scores, at the population level, University A will outrank University B.

However, in this simulation as in the THE survey, only a small fraction of students are sampled and these come disproportionately from those with more positive views of their institutional student experience, while being taken roughly equally across the range of views of disciplinary student experience. The response from each *sampled* student is given by a small black triangle in figure 6. The center of those sampled responses (marked with the large black triangle) represents the rating the institution receives from the sample who participate.

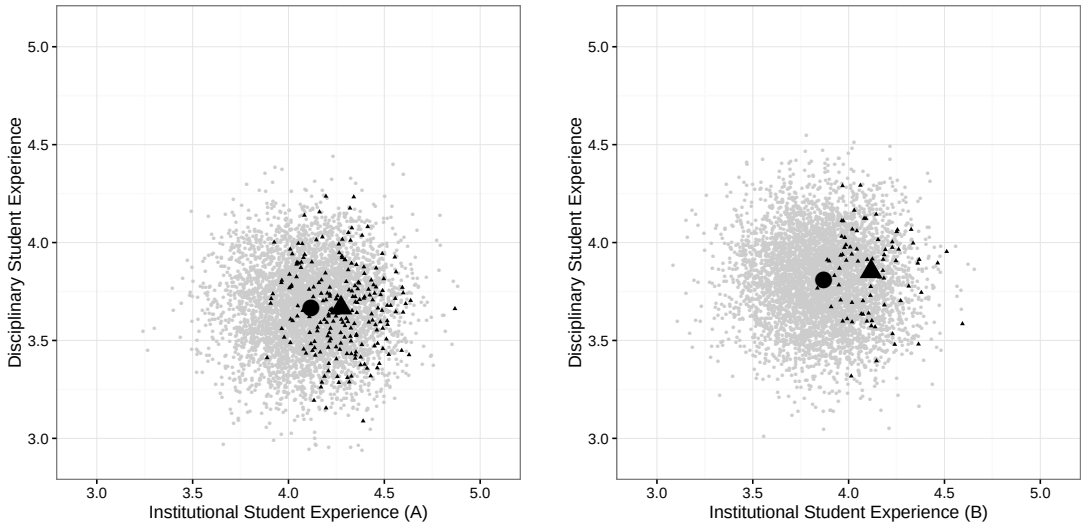


Figure 6: Two simulated universities. Small grey dots represent the responses from the whole population; small black triangles represent the response from the (biased) sample; the large black circle is the centre of the population responses and the large black triangle is the centre of the sampled responses.

The simulation highlights a number of things.

- The rating given by the sample for institutional student experience is higher than the rating from the population as a whole (that is, the large black triangle is shifted right of the black circle).
- There is little up or down shift between the population and sample centres, other than that due to ordinary sampling error. That is, the sample rating for disciplinary student experience is a more accurate estimate of the population rating.
- The right shift is larger for University B than University A.
- The sample for University B is smaller than the sample for University A.

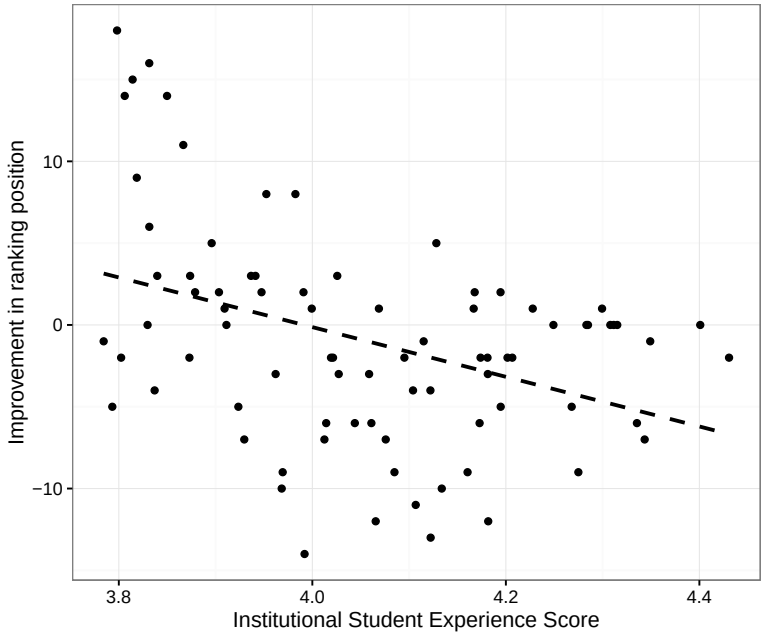


Figure 7: Simulation of 100 institutions with biased sampling

The disproportionate right shift for University B means that, when the two components are aggregated, University B outranks University A when we only look at the sampled responses. That is, the ranking reported by the simulated survey is the reverse of the one that would have be obtained if the the whole population had been sampled (or had the sample not been systematically biased).

To see the extent of this issue, one can repeat the simulation for 100 institutions and look at the extent of the change in ranking due to this systematic sampling bias. Figure 7 shows that those institutions which would have higher institutional disciplinary scores with random sampling tend to be disadvantaged by this bias.

7 Discussion and Conclusions

Surveys beget rankings. Yet compilers often give little thought to the ways in which barely distinguishable data points sitting in multiple dimensional spaces are reduced to uni-dimensional, definite rank orders. This is not for want of trying from statisticians and academic data analysts, as the concerns about rankings have been raised repeatedly for decades (Webster, 1981; Johnes, 1996; Altbach, 2006; Saisana, d'Hombres, & Saltelli, 2011).

Many of those concerns apply to league tables compiled from data from external sources, although some have noted their concerns with the highly subjective nature of tables compiled partially from opinions (e.g. Brooks, 2005). In the case explored here, the league table is compiled entirely from opinions.

It clearly suffers from some of the concerns noted in other league tables: it compares institutions which potentially have quite different missions; it aggregates responses with weightings which have little grounding (and which, in general, advantages institutions with traditional intakes); it occasionally modifies methods between years without strong justification yet treats league tables from different years as comparable and it fails to acknowledge

that very small, statistically insignificant differences in responses can be associated with very large differences in rankings.

These issues are compounded with the further problems of dimensionality and sampling bias outlined in this paper. The survey responses are not well modelled by a single dimension and both a statistical and theoretical perspective on the survey results suggest there are separate component scores measuring institutional and disciplinary student experience. In addition, there is a surprising, yet clear relationship between the sample size and the rating. This appears to come almost entirely from an underlying link between the institutional experience score and sample size. There may be a number of potential reasons for this: students who are enthusiastic about being part of a student community and who engage with sports and student representation may also be more likely to join the panel and complete the survey. It may also be that this survey is much easier to 'game': unlike a government survey (such as the NSS) there is no obvious penalty for overtly encouraging students to respond positively.

Whatever the cause, it is likely that the institutional student experience component will be an inflated estimate because the sample appears to systematically favour positive institutional responses. It is also likely that the disciplinary student experience component will be a much less precise estimate: while the respondents for a given institution probably share the same institutional-level facilities, they may come from many different disciplines and thus have much more varied views of disciplinary issues. That is, one dimension is measured more precisely but inaccurately and the other is measured more accurately but less precisely.

Importantly, when multidimensionality interacts with this systematic sampling bias, rank orderings can be reversed.

Goldstein (2014) states "I am not suggesting that league tables should never be published. Quantitative data that bear on performance are a useful tool for addressing the clear need for accountability from public (and other) institutions. When such data are reported publicly, however, their quality and reliability need to be displayed so that users of the data are not misled about what can be inferred. To withhold information about the uncertainty of rankings is to deprive users of information to which they are entitled." (p.398).

The evidence here suggests that we need more than information about the quality and reliability of the data: subtle but serious effects can occur from interactions between design and analysis (as here between panel design and uni-dimensional analysis). This implies that more considered analytic investigations need to take place before publishing survey results: declaring 'winners and losers' or 'risers and fallers' should be avoided, even if it does sell magazines.

References

Altbach, P. (2006). The dilemmas of ranking. *International Higher Education*(42), 2-3.

Attwood, R. (2009, 15th January). Happy to be here. *Times Higher Education*.

Blasi, B., Romagnosi, S., & Bonaccorsi, A. (2016). Playing the ranking game: media coverage of the evaluation of the quality of research in italy. *Higher Education*, *Online first*, 1–17.

Briggs, V. (2016). *Times Higher Student Experience Survey 2016.*

Brooks, R. (2005). Measuring university quality. *The Review of Higher Education*, *29*(1), 1–21.

Cheng, J. H., & Marsh, H. W. (2010). National student survey: are differences between universities and courses reliable and meaningful? *Oxford Review of Education*, *36*(6), 693–712.

Clarke, M. (2007). The impact of higher education rankings on student access, choice, and opportunity. *Higher Education in Europe*, *32*(1), 59–70.

Daraio, C., Bonaccorsi, A., & Simar, L. (2015). Rankings and university performance: A conditional multidimensional approach. *European Journal of Operational Research*, *244*(3), 918–930.

Dill, D. D., & Soo, M. (2005). Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems. *Higher Education*, *49*(4), 495–533.

Gibbons, S., Neumayer, E., & Perkins, R. (2015). Student satisfaction, league tables and university applications: Evidence from britain. *Economics of Education Review*, *48*, 148–164.

Gill, J. (2014, May). Competition drives excellent offerings. *Times Higher Education*, 3.

Goldstein, H. (2014). Using league table rankings in public policy formation: Statistical issues. *Annual Review of Statistics and Its Application*, *1*, 385–399.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 385–443.

Harvey, L. (2008). Rankings of higher education institutions: A critical review. *Quality in Higher Education*, *14*(3), 187-207.

Haunsperger, D. B. (2003). Aggregated statistical rankings are arbitrary. *Social Choice and Welfare*, *20*(2), 261–272.

Hazelkorn, E. (2008). Learning to live with league tables and ranking: The experience of institutional leaders. *Higher Education Policy*, *21*(2), 193–215.

Johnes, J. (1996). Performance assessment in higher education in Britain. *European Journal of Operational Research*, *89*(1), 18–33.

Longden, B. (2011). Ranking indicators and weights. In J. Shin, R. Toutkoushian, & U. Teichler (Eds.), *University rankings: Theoretical basis, methodology and impacts on global higher education* (pp. 73–104). Springer.

Messick, S. (1995). Standards of validity and the validity of standards in performance asessment. *Educational measurement: Issues and practice*, *14*(4), 5–8.

Porter, S. R., & Umbach, P. D. (2006). Student survey response rates across institutions: Why do they vary? *Research in Higher Education*, *47*(2), 229–247.

Proulx, R. (2007). Higher education ranking and leagues tables: lessons learned from benchmarking. *Higher Education in Europe*, *32*(1), 71–82.

Ramsden, P., & Callendar, C. (2014). *Review of the national student survey (Appendix A: Literature review)* (Tech. Rep.). NatCen.

Saisana, M., & d'Hombres, B. (2008). *Higher education rankings: Robustness issues and critical assessment* (Tech. Rep. No. 23487 EN). JRC Scientific and Technical Reports.

Saisana, M., d'Hombres, B., & Saltelli, A. (2011). Rickety numbers: Volatility of university rankings and policy implications. *Research policy*, *40*(1), 165–177.

Salmi, J., & Saroyan, A. (2007). League tables as policy instruments. *Higher Education Management and Policy*, *19*(2), 1–38.

Shin, J. C., & Toutkoushian, R. K. (2011). The past, present, and future of university rankings. In J. C. Shin, R. K. Toutkoushian, & U. Teichler (Eds.), *University rankings: Theoretical basis, methodology and impacts on global higher education* (pp. 1–16). Springer.

Soh, K. (2013). Times higher education 100 under 50 ranking: old wine in a new bottle? *Quality in Higher Education*, *19*(1), 111–121.

Usher, A., & Savino, M. (2007). A global survey of university ranking and league tables. *Higher Education in Europe*, *32*(1), 5–15.

van der Wende, M., & Don, W. (2009). Rankings and classifications: The need for a multidimensional approach. In F. A. van Vught (Ed.), *Mapping the higher education landscape* (pp. 71–86). Springer.

Webster, D. S. (1981). Advantages and disadvantages of methods of assessing quality. *Change: The Magazine of Higher Learning*, *13*(7), 20–24.

Williams, R., & de Rassenfosse, G. (2016). Pitfalls in aggregating performance measures in higher education. *Studies in Higher Education*, *41*(1), 51–62.

Yorke, M. (1997). A good league table guide? *Quality Assurance in Education*, *5*(2), 61–72.

Yorke, M., Orr, S., & Blair, B. (2014). Hit by a perfect storm? Art & design in the national student survey. *Studies in Higher Education*, *39*(10), 1788–1810.

YouthSight. (2015). *YouthSight panel book: Comprehensive summary of the YouthSight panel.* YouthSight: London.