How to use the National Student Survey (responsibly)

1. Introduction

The National Student Survey is a powerful force in UK higher education. It has a strong influence on how institutions are presented in the media, how they are judged by regulators and sector agencies, and on quality enhancement and assurance activities within institutions. At the same time as being widely derided as simplistic and misguided it has brought greater attention to the quality of learning and teaching (Buckley 2012), and even among institutional leaders there is a level of cognitive dissonance:

> "It is clear that higher education institutions are trying very hard to improve student satisfaction, even when they are sceptical about its meaning or educational consequences". (Gibbs 2012, p.14)

Despite its visibility in the media, its role in regulation and its influence within institutions it is poorly understood. It has been under-researched considering its power in the sector (Ashby et al 2011, Lenton 2015), and it is sobering to compare the limited amount of research on the NSS with the volume of publications around the National Survey of Student Engagement used in the US.[1] One consequence of the lack of research attention to the NSS is the simplistic use of the data. Firm conclusions are drawn from small differences in raw scores, without consideration of the statistical properties of the data. This is not a problem specific to the NSS, as evidence suggests that student feedback data is often over-interpreted (Abrami 2001, Boysen et al 2014). However it does reinforce the need for more sophisticated use of the NSS results:

> Disagreement about the validity and use of teaching evaluations is common but everyone can agree that improper interpretation of data from teaching evaluations should be avoided. (Boysen et al 2014, p.653)

2. Statistical significance and the NSS

One issue with the way that NSS results are used is to do with the representation of statistical significance. Statistical significance, in the context of a survey like the NSS, is a way of determining whether a particular difference in results is likely to be the result of chance, or represents a genuine and meaningful difference. If an institution's NSS score for a particular question changes from one year to the next, we will want to know whether that difference in score indicates a change in the underlying phenomenon (students' experiences) or is a result of random variation in the data. In sample surveys the concept is straightforward: significance testing is used to determine whether a difference could be due to the sample being unrepresentative of the population from which it is drawn. In a census survey like the NSS, the idea of significance testing is more complicated (and contentious) as the respondents are not a random sample of the population. Nevertheless, significance testing is a helpful way of determining how meaningful an observed difference is, and is appropriate on the basis that conclusions are drawn from NSS results about the experiences of potential students (see HEFCE 2014).

[1] An extensive but still only partial list is available at http://nsse.indiana.edu/html/featured_publications.cfm

Tests of statistical significance are often facilitated using the idea of a confidence interval, which is a range around the observed value in which the 'true' value can be expected to fall. Different levels of confidence can be used. A common level – and that used in the provision of confidence intervals for the NSS – is 95%. The interval then represents the range in which 95% of observed values would fall if the survey were repeated many times (with everything held constant). So for an NSS percentage agreement score of 70%, confidence intervals ranging from 60% to 80% indicate that if the survey were to be run repeatedly (in the same circumstances) 95% of the time between 60% and 80% of the students would register agreement. Once the desired level of confidence has been chosen (95% in our case) the width of the confidence interval is effected by the number of responses and the variability in those responses.

Statistical significance is a key issue for the NSS. Comparisons between institutions involve large number of responses, and confidence intervals are correspondingly small. However, subject of study seems to affect students' responses much more strongly than the institution in which they study (Surridge 2009), meaning that an overall institutional score is strongly influenced simply by the proportions of students in different subjects. It is also the case that for the key purposes of exploring NSS results, subject rather than institution is a far more appropriate level of analysis: prospective students will be more interested in data that relates specifically to the course they wish to pursue, and the organisational structures of higher education institutions mean that the relevant focus for enhancement efforts and accountability is likely to be department level. However, at subject level the number of responses can become problematically small and the confidence intervals correspondingly large:

> "Courses explain more variance in NSS responses than universities. However, because the number of students in each course within a given university is so much smaller, these ratings lack reliability and few differ significantly from the grand mean." (Cheng and Marsh 2010, p.707)

The challenge of reporting data in a way that is at a meaningful – i.e. subject – level and still reliable was noted by HEFCE at the point the NSS was developed (HEFCE 2004b).

The importance of statistical significance for the presentation of NSS results has been noted numerous times over many years. The consultation on the introduction of the NSS in 2004 indicated that 95% confidence intervals ("to indicate the statistical reliance that can be placed on the score", p.12) would be included on the public presentation of the data – then the Teaching Quality Information website (HEFCE 2004a). While confidence intervals have been available for the publicly-downloadable spreadsheets since 2007, the primary mechanism for the public presentation of the results (TQI and then Unistats) has not included information about statistical significance. The consultation on the revisions to the NSS in 2015 again highlighted the need for better explanation of the effect of sample sizes on comparisons of the data (HEFCE 2015). Researchers have also made the point:

> "Rank-order differences between universities and courses like those used to construct league tables typically reflect substantial amounts of random

> error… Indeed, at the university level, there are relatively few universities that differ significantly from the mean across all universities and, at the course level, there is even a smaller portion of differences that are statistically significant. This suggests the inappropriateness of these ratings for the construction of league tables… When results of the NSS are presented or used for any of their intended purposes, this substantial error variance identified here should be emphasised appropriately. Thus, for example, error bars (or confidence intervals)…would make clear that differences between individual universities are mostly unreliable." (Cheng and Marsh 2010, pp.708-9)

In 2012 HEFCE introduced institutional benchmarks, apparently in response to these kinds of concerns. The benchmarked scores compare an institution's score for overall satisfaction to the score that it 'should' have got, when the mix of students at that institution is taken into account; whether the actual score is statistically significantly better or worse than the benchmark is also reported. This idea has been transferred to the Teaching Excellence Framework, where an institution's 'flags' are determined by whether the differences between their NSS results and their benchmarked score are statistically significant.

These are developments in how NSS results are presented to a mass audience. For institutions themselves – and anyone else interested in looking deeper – since 2007 HEFCE (and now the Office for Students) have made 95% confidence intervals for NSS data publicly available via their website. Until 2015 these were of limited institutional use, as the publication threshold for publicly available data was higher than for the data provided confidentially to institutions: 23 responses rather than 10. However in 2015 the publication threshold for the publicly available data was reduced to 10, bringing the two datasets in line. Table 1 is an example from the publicly available dataset (for Anglia Ruskin University for the subject of Drama) for the most fine-grained level of the Joint Academic Coding System (JACS), JACS 3, containing 108 subjects.[2]

Table 1: Example of publicly available NSS data

			%Agree			
Institution	Subject	Question Number	Confidence interval - min	Actual value	Confidence interval - max	Response
Anglia Ruskin University	(098) Drama	Q01	81%	93%	98%	62
Anglia Ruskin University	(098) Drama	Q02	64%	79%	89%	62
Anglia Ruskin University	(098) Drama	Q03	57%	72%	83%	62
Anglia Ruskin University	(098) Drama	Q04	37%	52%	66%	62
Anglia Ruskin University	(098) Drama	Q05	58%	73%	84%	62
Anglia Ruskin University	(098) Drama	Q06	55%	70%	82%	62
Anglia Ruskin University	(098) Drama	Q07	51%	66%	79%	62

[2] NSS data with confidence intervals are available to download at https://www.officeforstudents.org.uk/advice-and-guidance/student-information-and-data/national-student-survey-nss/get-the-nss-data/

Anglia Ruskin University	(098) Drama	Q08	48%	63%	76%	62
Anglia Ruskin University	(098) Drama	Q09	34%	49%	64%	62
Anglia Ruskin University	(098) Drama	Q10	57%	72%	83%	62
Anglia Ruskin University	(098) Drama	Q11	55%	70%	82%	62
Anglia Ruskin University	(098) Drama	Q12	71%	85%	93%	62
Anglia Ruskin University	(098) Drama	Q13	58%	73%	84%	62
Anglia Ruskin University	(098) Drama	Q14	59%	74%	85%	61
Anglia Ruskin University	(098) Drama	Q15	57%	72%	83%	62
Anglia Ruskin University	(098) Drama	Q16	67%	81%	90%	62
Anglia Ruskin University	(098) Drama	Q17	41%	57%	71%	60
Anglia Ruskin University	(098) Drama	Q18	60%	75%	86%	62
Anglia Ruskin University	(098) Drama	Q19	71%	85%	93%	62
Anglia Ruskin University	(098) Drama	Q20	61%	76%	86%	62
Anglia Ruskin University	(098) Drama	Q21	41%	56%	70%	62
Anglia Ruskin University	(098) Drama	Q22	68%	83%	91%	61
Anglia Ruskin University	(098) Drama	Q23	59%	74%	85%	61
Anglia Ruskin University	(098) Drama	Q24	53%	68%	81%	61
Anglia Ruskin University	(098) Drama	Q25	41%	56%	70%	61
Anglia Ruskin University	(098) Drama	Q26	42%	57%	71%	61
Anglia Ruskin University	(098) Drama	Q27	54%	69%	81%	61

Table 1 shows that an institutional score for Q1 of 93% based on responses from 62 students, can have confidence intervals ranging from 81% to 98%. So we can be 95% confident that the true value falls somewhere within that range. And 62 is the average number of responses for an institution at JACS 3 subject level, so many institutional scores will have much wider confidence intervals. The average width of the confidence intervals at JACS 3 subject and question level (from the minimum to the maximum confidence interval) is 29%, e.g. for a score of 80%, the true value could be expected to fall anywhere between – on average – 65% and 94%. This highlights the imprudence of relying on the raw scores.

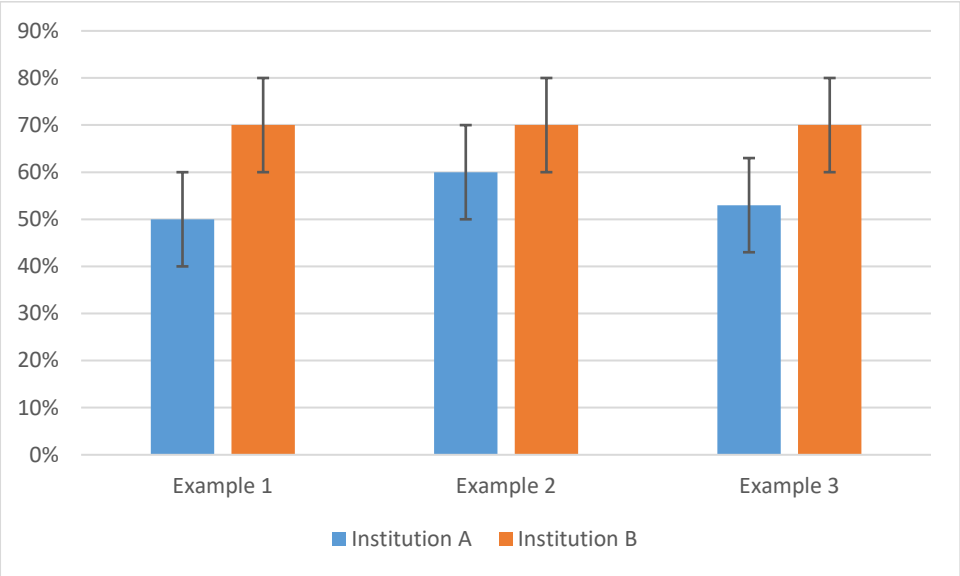
3. Ways of using confidence intervals on NSS data

The confidence intervals allow a range of comparisons that take into account the random variation in the data, using the evaluation of overlap between confidence intervals. For 95% confidence intervals, if they don't overlap then the difference is statistically significant. However it isn't the case that if they overlap at all then the difference is not statistically significant. That confidence intervals can overlap without the loss of statistical significance fact is apparently not widely known even among those who are familiar with statistics (Krzywinski and Altman 2013).

An approximate rule of eye, developed by Cumming and Finch (2005), is that if the overlap of the confidence intervals for two values is no more than half of the average length of the confidence intervals then the difference can be taken to be statistically significant. Figure 1 shows three example comparisons between institutional scores. In Example 1, there is no overlap between the confidence intervals, indicating statistical significance. In Example 2,

the overlap between the confidence intervals is large and we can conclude that the difference is not statistically significant. In Example 3, there is an overlap between the confidence intervals, but the overlap is less than half of the average confidence interval, indicating statistical significance. [3]

Figure 1: Examples of error bar overlap



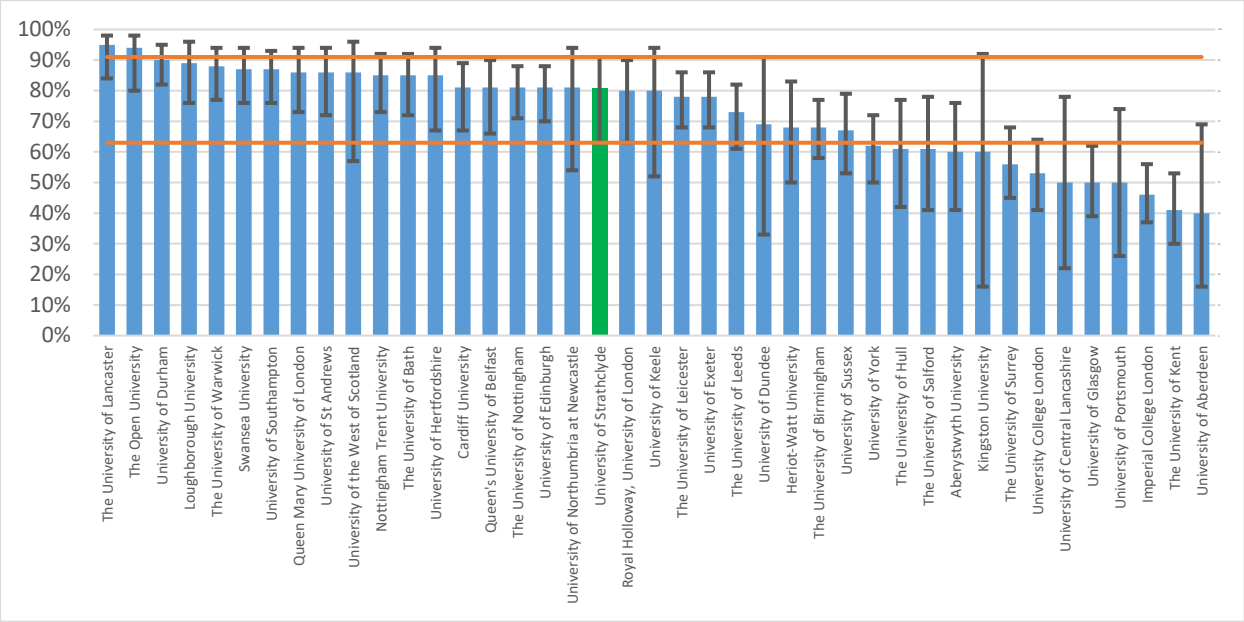
The evaluation of the overlap of confidence intervals permits three kinds of important comparisons: i) comparisons between institutions; ii) comparisons with the subject average; and iii) comparisons over time.

3.1 Comparisons between institutions

The presentation of NSS results within institutions is often based on comparisons with other institutions. The inspection of confidence intervals allows this to be done in a more meaningful way. Summary measures (such as institutional heatmaps) cannot be developed using the rule of eye described above, as it is not intended to yield precise calculations, but to be "easily remembered, pragmatically useful guidance for anyone inspecting a figure that presents data" (Cumming 2009, p.207). In lieu of summary measures, inspection of institutional scores with confidence intervals is a helpful, if time-consuming, way of making sense of comparisons of a range of institutions' NSS results. Figure 2 shows an example of this kind of chart. It shows institutions' results for Q10 for the subject of Physics and Astronomy. With an average of 61 responses per institution, this is a representative subject. Many subjects have smaller numbers of responses, and for them the confidence intervals will tend to be larger. Also included on the chart are horizontal lines representing the upper and lower limits of the confidence intervals for the NSS score for Strathclyde, for ease of visual inspection.

Figure 2: Example of institutional comparison (Q10 for Physics and Astronomy)

[3] For more on interpretation of the interpretation of confidence interval overlap, including the rule of eye described here, see Cumming (2009) and Krzywinski and Altman (2013).



Visual inspection of Figure 2 indicates that few of the institutions differ meaningfully in their NSS results for this subject and this question. For example, the result for Strathclyde has confidence intervals that fail to overlap with those from only three other institutions (Glasgow, Imperial and Kent). There are three further institutions for which the overlap is less then half of the average interval (Surrey, UCL and Aberdeen). So visual inspection of Figure 2 indicates that for Q10 for the subject of Physics and Astronomy, Strathclyde is statistically significantly different to only six other institutions, of the 40 whose data is available. And despite being ranked mid-table by raw score (19th of 41), Strathclyde does not have a statistically significantly lower score than any other institution.

3.2 Institutional score vs subject average

Sector averages are provided at subject level, which in conjunction with confidence intervals for institutional results allows institutions to determine whether they differ meaningfully from the average performance of the sector.[4] While HEFCE/OfS do not provide confidence intervals for the sector averages, the number of responses that contribute to a sector average at subject level is sufficiently large that we can assume that the confidence intervals are effectively zero. Therefore, if the sector average lies outside the confidence intervals for an institutional score, we can infer that there is no overlap in confidence intervals and so the difference is statistically significant.

Unlike the comparison between institutions in Figure 1, this form of analysis does provide for a clear and accessible summary, in the form of a 'heatmap'. These are often created from raw NSS scores or using measures such as quartile values, but it is possible to create heatmaps that take statistical significance into account. This way of presenting the data also has the benefit that it does not depend on individuals being able to interpret information about statistical significance (such as confidence intervals); evidence suggests that even

[4] For some reason sector averages are not made publicly available via the HEFCE or OfS websites, instead they are included with the NSS data provided confidentially to institutions.

those with statistical training tend to overlook statistical information on student feedback data (Boysen 2017). Figures 3 and 4 below show heatmaps from two institutions, one that is in the top 10 institutions for overall satisfaction, and another that is in the bottom 10 institutions for overall satisfaction. These institutions would therefore be seen as 'high-performing' and 'low-performing', respectively, in the NSS. The colours in the heatmaps indicate where institutional results differ meaningfully (either positively or negatively) from the sector average for that subject: red cells indicate results that are statistically significantly lower than the sector average (for that subject and question); green cells indicate results that are statistically significantly higher than the sector average; and grey cells indicate results that are not statistically significantly different from the sector average.

Figure 3: Institutional heatmap, 'high performing' institution

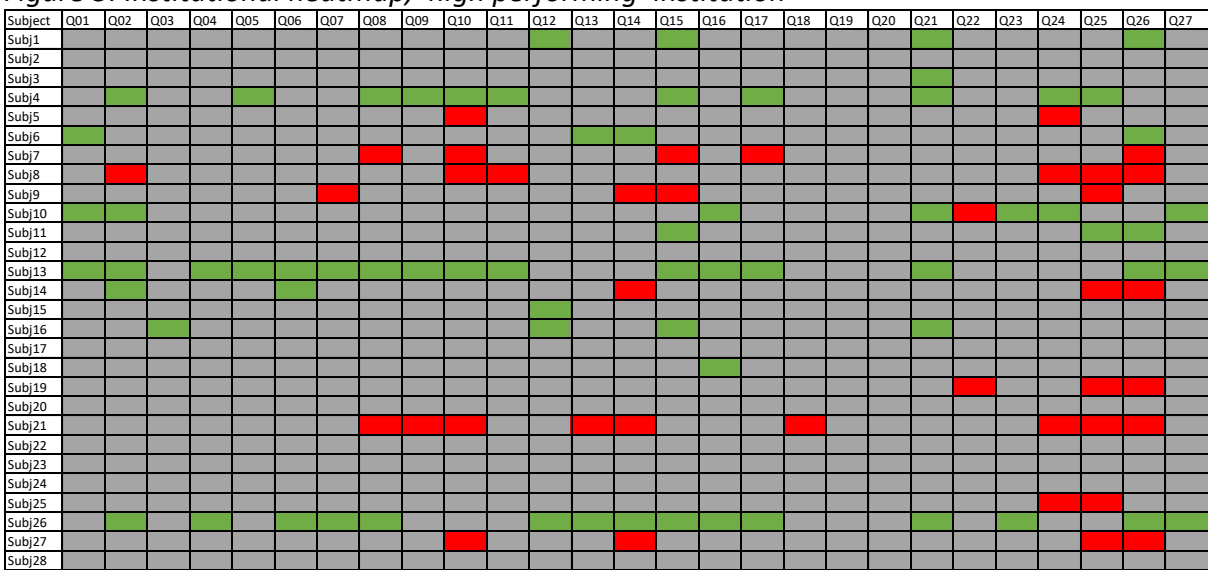
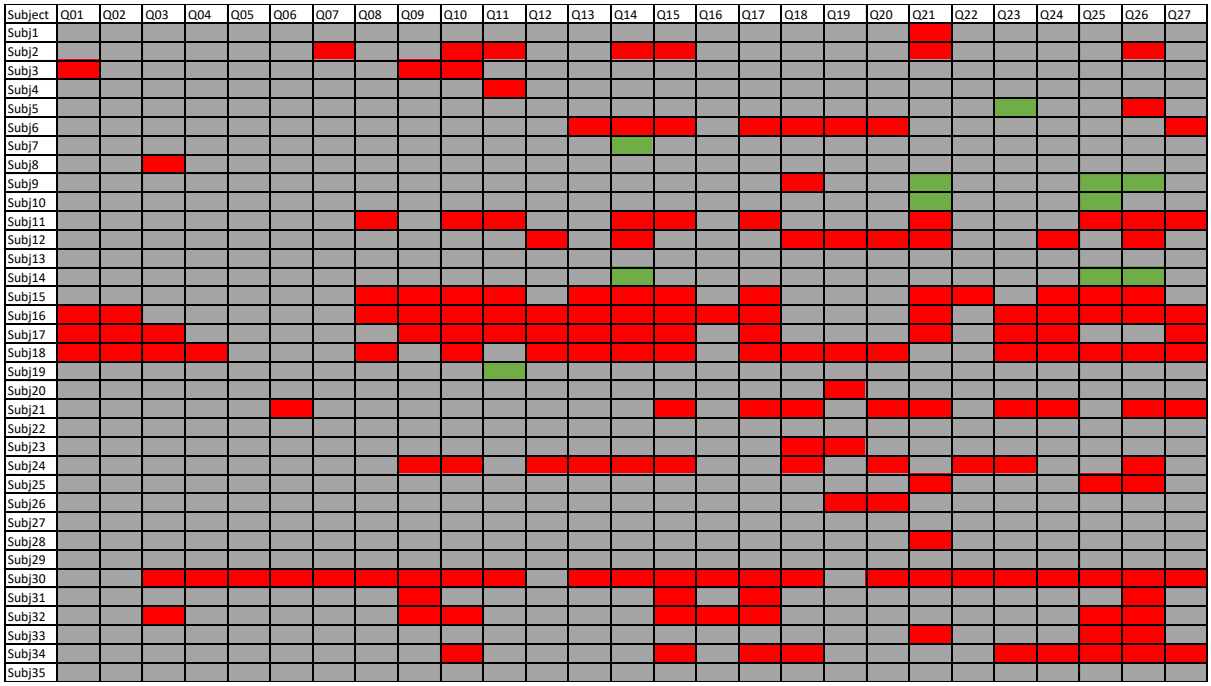


Figure 4: Institutional heatmap, 'low performing' institution



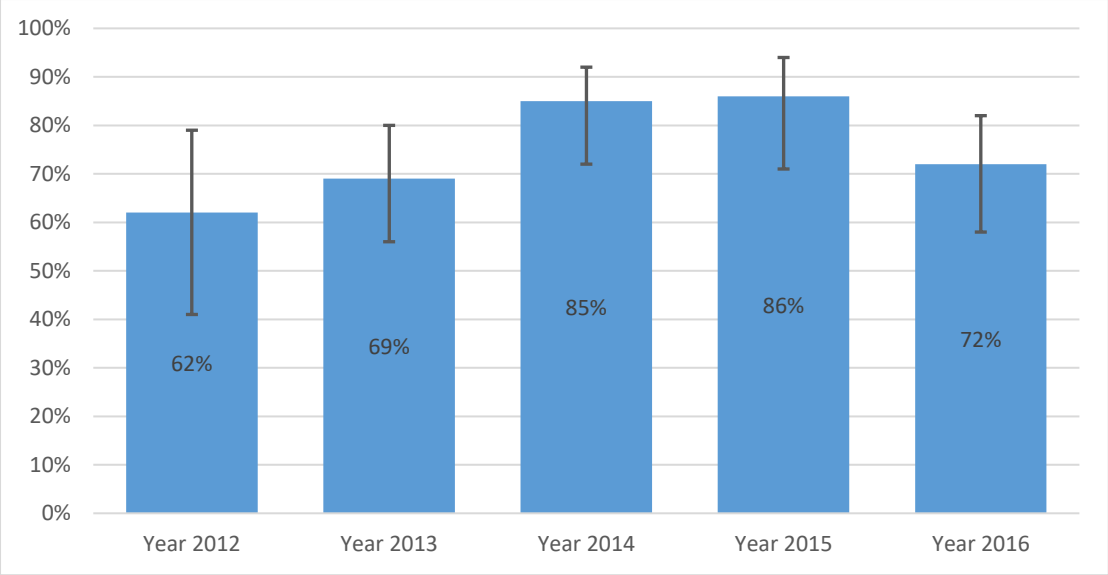
In both Figures 3 and 4, the vast majority of differences between the institution's score and the sector average are not statistically significant. Heatmaps based on raw scores (e.g. quartiles) often suggest that for every subject and every question, an institution has performed 'well' or 'badly'. However it is highly likely that for most subjects and most questions, an institution's NSS results do not differ meaningfully from the sector average. One of these institutions would be held up as a very high performing institution, while the other would be taken to have significant challenges, but the heatmaps show that the differences in performance in the NSS are in fact relatively small. For around 80% of the results at subject and question level, neither institution differs meaningfully from the sector average.

3.3 Change in institutional score

Another prominent measure within institutions is the change in scores over time. The use of raw scores can often give the impression of substantial fluctuation between years, with academic staff struggling to discern reasons for changes in students' perceptions. The use of confidence intervals reveals that many of these changes between years are not meaningful. Figure 5 below shows an example of institutional scores for Q2 from a particular subject over a number of years. For this institution and subject, there were around 60 responses each year, making the size of the confidence intervals reasonably representative. Despite considerable change in the raw scores between years (e.g. 69% in 2013 to 85% in 2014) none of the changes between years reaches the level of statistical significance, using the rule of eye described above.[5] Effort by staff to explain these observed year-on-year changes is likely to be a waste of time.

Figure 5: Example of change in scores 2012-2016

[5] The NSS questionnaire changes in 2017, preventing comparisons with that year.



4. Conclusion

The NSS has considerable power to celebrate or shame departments and by implication individual teaching staff. NSS results have become the primary way of judging the quality of teaching and the student experience. However, there are significant limitations to the reliability of comparisons at subject level due to the small number of responses, and this has in fact been acknowledged by those who run the survey. In the consultation on the reduction of the publication threshold from 23 responses to 10, HEFCE provided analysis indicating that the proportion of overlap of confidence intervals at subject level - using 21 subject groups, a broader grain of categorisation than standardly used within institutions – ranged between 72% and 97% for a threshold of 23 responses, and between 73% and 98% for a threshold of 10 responses (HEFCE 2014).[6]

While the use of NSS results for league tables is largely outside the influence of those working in institutions, the limitations of the data can be addressed to some extent within institutions using the confidence intervals made publicly available. At present, the use of raw scores often gives a specious meaningfulness to the comparisons presented to academic staff, and a "common concern is that administrators do not have adequate expertise on the use and interpretation of…teaching evaluation results. For example, administrators might believe that means can be interpreted to the third decimal, that all fluctuations up and down are interpretable trends, or that means falling below a specific standard are always indications of poor teaching" (Boysen et al 2014, p.642).

In this paper I have described three ways in which the confidence intervals can be used to present NSS results in a way that takes the small number of responses at subject level into account. These methods use the idea of statistical significance to highlight comparisons that are meaningful. By doing so, they communicate the limitations of the NSS data, and allow more evidence-based consideration of institutional performance in learning and teaching.

[6] The small difference caused by the reduction of the threshold was used as justification for the change. It was incidental that it showed that even with the higher threshold the vast majority of comparisons between institutions at subject level are not statistically significant.

References

Abrami, P. (2001) 'Improving judgements about teaching effectiveness using teacher rating forms', *New Directions for Institutional Research* 109: 59-87

Ashby, A., Richardson, J. and Woodley, A. (2011) 'National student feedback surveys in distance education: An investigation at the UK Open University', *Open Learning* 26(1): 5-25

Boysen, G. (2017) 'Statistical knowledge and the over-interpretation of student evaluations of teaching', *Assessment and Evaluation in Higher Education* 42(7): 1095-1102

Boysen, G., Kelly, T., Raesly, H. and Casner, R. (2014) 'The (mis)interpretation of teaching evaluations by college faculty and administrators', *Assessment and Evaluation in Higher Education* 39(6): 641-656

Buckley, A. (2012) *Making it count: Reflecting on the National Student Survey in the process of enhancement* (York, Higher Education Academy)

Cheng, J. and Marsh, H. (2010) 'National Student Survey: Are differences between universities and courses reliable and meaningful?' *Oxford Review of Education* 36(6): 693-712

Cumming, G. (2009) 'Inference by eye: Reading the overlap of independent confidence intervals', *Statistics in Medicine* 28: 205-220

Cumming, G. and Finch, S. (2005) 'Inference by eye: Confidence intervals and how to read pictures of data', *American Psychologist* 60(2): 170-180

Gibbs, G. (2012) *Implications of 'Dimensions of Quality' in a market environment* (York, Higher Education Academy)

HEFCE (2004a) *National Student Survey 2005: Consultation* (Bristol, Higher Education Funding Council for England)

HEFCE (2004b) *National Student Survey 2005: Outcomes of consultation and guidance on next steps* (Bristol, Higher Education Funding Council for England)

HEFCE (2014) *Data publication thresholds and aggregation on Unistats: Consultation on thresholds and subject groupings for data on Unistats and the National Student Survey* (Bristol, Higher Education Funding Council for England)

HEFCE (2015) *Review of information about learning and teaching, and the student experience: Consultation on changes to the National Student Survey, Unistats and information provided by institutions* (Bristol, Higher Education Funding Council for England)

Krzywinski, M. and Altman, N. (2013) 'Points of significance: Error bars', *Nature Methods* 10(1): 921-922

Lenton, P. (2015) 'Determining student satisfaction: An economic analysis of the National Student Survey', *Economics of Education Review* 47: 118-127

Surridge, P. (2009) *The National Student Survey three years on: What have we learned?* (York, Higher Education Academy)