Professor John MacInnes, Professor of Sociology, University of Edinburgh.
Strategic Advisor on Quantitative Skills, The British Academy.

Personal commentary submitted along with the British Academy's response to the Department for Education's consultation on the Teaching Excellence and Student Outcomes Framework: Subject-level.

This commentary is on the methodology of the Subject Level TEF as set out in the Department for Education's *TEF Subject level Consultation Technical document*. I have also used the original TEF Framework specification to which it refers, to the guidance on benchmarking available on the old HEFCE website, to the ONS review of data sources for the TEF, and to brief email correspondence with the *Office for Students*.

At some points the specification is ambiguous, and so what follows is based on my best inference as to their interpretation. Much of what follows is unavoidably technical, however it would perhaps be a mistake to focus on the details of the technical specification if by dint of that, we lose sight of the essence of the problem facing *any* approach to measuring quality at subject level within providers. This is that students typically take courses on which the numbers of students in any year are far too small to allow any meaningful analysis to take place. Conversely, the aggregation of students into groups large enough to make meaningful *reliable statistical* analysis possible (and it must be stressed, such groups need to be surprisingly large) debases the *validity* of the analysis by treating disparate groups of students with a variety of educational experiences and studying different subjects, as if they were in fact homogeneous. This dilemma is rooted in the laws of mathematics, it cannot be overcome. It is the Scylla and Charybdis of the subject level TEF.

Students typically undertake their study in small groups. Students make choices about courses or degree programmes.UK universities currently offer 37,148 different programmes (UCAS data accessed 3rd May 2018). There were 548,425 first year students on UG programmes in 2016/7, giving an average annual cohort of just under **15** students per programme. This is an average, and it is probably the case that a substantial number of degree programmes do not recruit students every year. However, few degree programmes have annual cohorts of students reaching three figures. I am not aware of any data on the

distribution of the size of degree programmes in the public domain. This would be useful information.

While the continuation, attitudes and performance of groups of 15 students could be measured, the results would have no value. Even if students *within* a course tended to take a common view of it, and experience common outcomes after it, and even if these common views and outcomes contrasted substantially *between* different courses, we would still find that random fluctuation noise would swamp the signal from any but the very strongest contrasts.

Some sense of this can be gleaned from the following table, which gives the approximate 95% confidence intervals for the proportion 0.5 obtained from samples of different sizes. A 95% CI gives the range of estimates for the parameter we are trying to measure, that we'd be likely to obtain in our 95% most successful attempts to measure it. It can be seen that the CIs for small numbers of observations, including the average cohort size of a degree programme, are so wide as to preclude drawing any usable conclusion.

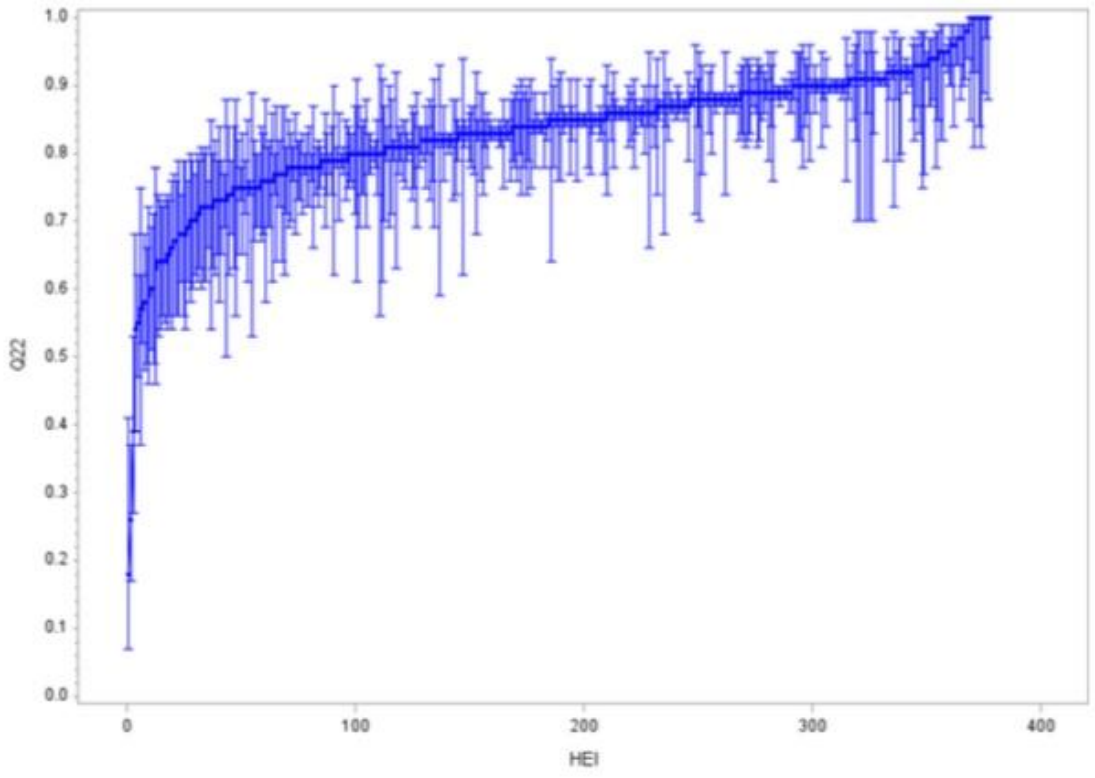
N responses	CI lower limit	CI upper limit
10	0.19	0.81
15	0.25	0.75
50	0.36	0.64
100	0.40	0.60
1,000	0.47	0.53
10,000	0.49	0.51

So far as I am aware, no technical review has produced confidence intervals for NSS data below that of provider level and for all students, or for DHLE data at any level. This is important as confidence intervals allow us to estimate how confident we can be that any difference we might observe between two measurements is one that is caused by a real difference in the characteristic measured, and not merely some artefact of the measuring process, including random variation in its reliability.

Until such confidence intervals have been published for subjects within providers, the quality of TEF subject level measure cannot be assessed. One of the reasons for this is that, in contrast to the hypothetical situation drawn above, previous analysis of NSS data has shown that the variance of student opinion and behaviour within the same department and provider is large, while that between departments and providers is small. Marsh and Cheung (2008) found that most of the variance in the results occurs at individual level, with only around 10% being attributable to some combination of provider and department

The result of this situation is illustrated by the only published CIs available, in the ONS technical report on the provider level TEF data sources (Figure 1, ONS 2016, reproduced below). Even though the unit of measurement here is whole providers, so that the number of responses is much larger than we would encounter at subject level the between provider variance is so modest that most providers cannot be differentiated from each other. The vertical bars represent 95% CIs. Those of all but the very best providers overlap with all but the very worst.

Figure 1: Percentage of students with positive response to Q22 – Overall satisfactio



As the ONS commented: 'For the main body of institutions in the middle of the graph, there does not appear to be any significant difference in the outcomes. It may be possible to separate institutions at the extremes of the distribution, those institutions with particularly low and particularly high proportions. However, *care would be needed in determining any such thresholds, to ensure the differences are significant*. A relatively similar distribution is seen for other questions on the survey' (emphasis added).

The ONS review team had access to results at provider level broken down by a single classification (e.g. sex, ethnicity) and concluded that 'it is likely that comparisons of raw data between institutions at this level would not be statistically significant. Whether this would also be an issue for the benchmarking approach would need to be examined'. Breaking down data by a single categorical variable (for example, by age or sex) generally gives large enough sample sizes for general statistical analysis (*but not for comparisons between institutions*), but breaking down data by more than one variable leads to insufficiently large sample sizes.

It can easily be seen that if sample size is a problem when analysing *provider* level data, when most universities have several thousand students, it will be greater when analysing *subject* level data.

The TEF subject level specification proposes two ways to work round this problem. It will not use courses or degree programmes but aggregate students by 35 CAH2 subjects and by 7 'subject groupings'. The former will produce 'subject instances' that will be the basis of most analysis. The specification gives no information on the size distribution of these instances, only an estimate of their number (about 4,500). Without this information the statistical viability of the subject level TEF cannot be assessed. How homogeneous will the students in each 'subject instance' be? It would be helpful for DfE to release NSS or DHLE data by subject instance to allow some analysis of this.

The specification asserts that the 35 CAH2 subject headings and the 7 broad subject groupings 'are likely to have similar teaching practices, teaching quality and student outcomes.' This is unlikely to be the case. There is little reason to expect teaching in Computing or Engineering to be similar, or Maths and Agriculture, or Architecture and Politics or Archaeology and French language. Yet this is what the specification claims. It *must* make this claim, for without this aggregation, no robust statistical analysis can be done.

The essence of the problem with the subject level TEF is that the unit that is of interest to most students is the degree course, or the subject area or department teaching it. This is also typically the lowest unit through which university governance and compliance processes operate and for good reason: the demands of teaching organisation, delivery and assessment are typically subject specific. That is why the remit of external examiners, for example, are usually specific to a department or degree programme, not spread across disparate units. However, they are typically far too small to produce meaningful comparisons of any but the crudest sort. The statistical noise of variance in student attitudes would swamp any but the very strongest signal.

Course instances and broad subject groupings produce aggregations of student experience and outcomes that are large enough to produce statistically meaningful results in which the relative size of confidence intervals is reduced. But the cost of statistical reliability is a consequent loss of validity, insofar as these groupings no longer correspond to choices that prospective students could make or to governance structures or quality assurance mechanisms that universities could operate. No actual student can experience, nor university manage, and quality assure these essentially statistical aggregates.

Benchmarking

Because of the difficulty of producing comparisons between small units that reach statistically significance, the TEF proposes to adopt the benchmarking procedure proposed by Draper and Gittoes (2004). In such an approach the raw data for individual students in a subject group is first adjusted to give these students the characteristics typical of all the students in that subject in the whole country. This adjusts for e.g. courses with higher or lower numbers of mature students, minority ethnic students and so on. Subject instances are then compared to the average for all sector subject instances (the 'Initial Hypothesis'), and those outside the confidence interval for this average are flagged. This procedure is useful because it reduces the relative size of confidence interval by comparing subject instances not with others, but with what could be thought of as a standardised subject instance.

There are, however, three key issues with benchmarking. First, it is not transparent. The ONS review of data sources called for an independent review of benchmarking, and in particular of the procedure for producing standard errors and the assumptions it makes. Such a review,

initially scheduled for 2017, has yet to start. The statistical properties of the benchmarks proposed are, to say the least, not well understood, especially where multiple post hoc comparisons are being made.

Second, when it takes place, the review needs to produce an account of the benchmarking process that can be understood by stakeholders and, above all, by those responsible for conducting the subject level TEF. Such an account is unavailable. Without it, results based upon benchmarking are likely to be abused in the same way as previous TEF statistics, for example, the production of leagues tables that erroneously assume a degree of accuracy that NSS scores cannot provide. Unless reviewers understand benchmarks, they are unlikely to use them properly. It is not clear how reviewers interpreted them in the pilots.

Third, a clearer account has to be given of the minimum number of student responses that will be permitted to be used to produce any TEF score. At times, the subject level specification creates the impression that it has been produced by authors who have not grasped the dramatic impact of small numbers on the ability to produce meaningful statistics. The specification (section 8.2) sees the problem of small cohort size as proportionality of effort, not small sample size. At times, when discussing procedures for NSS student boycotts, the specification appears to imply that Ns as low as 5 or 10 responses are contemplated. I assume that this must either be a drafting error, or a misapprehension on my part. Nothing can possibly be gleaned from such numbers.