ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

Response to the Department for Business Innovation and Skills' Technical Consultation (year 2) on the Teaching Excellence Framework

Organisation: The Royal Statistical Society (RSS)

Please tick the box that best describes you as a respondent to this consultation:

☒	Other (please describe)

Learned society and professional body

Question 1 (Chapter 1)
Do you agree with the criteria proposed in Figure 4?

☐Yes ☒ No ☐ Not sure

Please outline your reasons and suggest any alternatives or additions.

It is concerning that Figure 4 equates student satisfaction as measured by the National Student Survey (NSS) with Teaching Quality. We are not aware that there is any evidence of a statistical association between the two concepts.

What some research does show is that there is no reliable association between the two. [1]

It may therefore be the case that innovative forms of teaching, teaching that challenges pre-conceptions or forces students to engage in new ways, or is in any way non-standard, tends to attract low student satisfaction ratings and yet, these teaching methods are often highly successful in boosting student learning.

[1] http://www.tandfonline.com/doi/abs/10.1080/13562510601102131 With the purpose of highlighting the validity and use of student evaluations of teaching (SETs), this article analyzes student grades and student evaluations of teaching performance along with nine other independent variables in the Spanish program of a major university. Data analyzed for this project represents four years of teaching and includes the evaluations and grades of 18,175 students. The most significant findings of the study are the following: There is a moderate correlation between low grades and low evaluations, but no correlation between high grades and high evaluations when all cases are considered together. Analysis and interpretation of the data suggests that SETs should not be used to compare language instructors. In addition, since the relationship between student evaluations and the actual merits of teaching performance has not been clearly identified, numerical values of those evaluations should not be used in critical personnel decisions such as retention, tenure and promotion of faculty, unless they are properly interpreted within a sound theory of teaching effectiveness.

This paper relates to languages – but there are other papers to this effect also. See: Measuring Teaching Effectiveness: Correspondence Between Students' Evaluations of Teaching and Different Measures of Student Learning http://link.springer.com/article/10.1007%2Fs11162-012-9260-9

Neither the proposed TEF nor the NSS will be able to recognise the absolute level of challenge that different programmes provide. Two HE institutions might offer programmes with the same name but containing different content and differing amounts of content. However, the NSS only seeks to ascertain satisfaction within specific programmes within institutions. Anecdotally, we have heard of institutions explicitly `dumbing down' programmes so as to result in higher NSS scores. A new TEF needs to recognise this and mitigate against it. One goal of higher education is to produce highly educated people of use to the society of the future and the NSS inadvertently encourages the opposite.

We are keen on criteria that reward innovation and effective outcomes for students from disadvantaged backgrounds. We believe that more statistical research is needed to ascertain what are the key factors for improvements in this area.

It is not clear that it is possible to discriminate the vast majority of HE institutions on the overall NSS satisfaction scores, let alone when they are broken down into smaller subgroups. Such a conclusion is supported by Figure 1 of "Teaching Excellence Framework: Review of Data Sources --- Interim Report" prepared by the Office for National Statistics' Methodology Advisory Service, February 2016 [2] and the following text: "Confidence intervals are not available for breakdowns of NSS data by sub-groups such as ethnicity and socio-economic classification" and "comparisons of raw data between institutions at this level would not be statistically significant".

Further comments on the criteria will be made below in answers to specific questions, notably on those relating to employment metrics.

Question 2 (Chapter 3)

A) How should we include a highly skilled employment metric as part of the TEF?

The inclusion of a highly skilled employment metric is premature for two reasons. First, there is a lack of evidence on whether employment outcomes are valid indicators of teaching quality. Second, there are doubts about whether the DLHE survey in its current form is a suitable data source; this is of particular concern in light of the ongoing review of DLHE and plans for the availability of the new Longitudinal Educational Outcomes dataset. We elaborate on these points below.

(i) *Validity of employment metrics as indicators of student outcomes and learning gain*. The inclusion of employment outcomes as part of TEF suggests that there is a causal link between the quality of teaching and a graduate's employment circumstances at an arbitrary date following completion of their degree course. The time to finding employment, and the type of job obtained, depends on individual characteristics that may not be adequately captured by the factors proposed for use in benchmarking. Further research is needed to assess the evidence for a link between employment outcomes and teaching quality. This should include an analysis of the extent of correlation between the proposed employment

[2] ONS Methodology Advisory Service (2016) *Teaching Excellence Framework: Review of Data Sources - Interim Report* (PDF) Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/523291/bis-16-269-teaching-excellence-framework-review-of-data-sources-interim-report.pdf

metrics and other more widely accepted measures of teaching quality, after accounting for student characteristics.

(ii) *Data sources*. As acknowledged in Annex D, it may take longer than 6 months for graduates to find employment. This is a particular issue when considering *type* of job (versus *any* job) as it takes significantly longer to establish a career in some professions than others. The proposed mitigation of allowing institutions to "provide longer-term destination data in their submission if they feel that this demonstrates a better success rate" is unsatisfactory as such data will not be externally validated and will not be comparable across institutions.

Before introducing a 'highly skilled' metric, research is needed on the time taken to attain highly skilled employment. Questions to consider might include the proportions in highly skilled employment after 6, 12 and 18 months, and the factors associated with longer times to entry into highly skilled employment. This research would be based on data sources providing longer-term outcomes such as the Longitudinal Educational Outcomes dataset and the longitudinal DLHE. (Although we agree that the longitudinal DLHE is unsuitable for deriving TEF metrics, it may provide useful information for research purposes.) If a 6-month window is found to be insufficient, the publication of 6-month employment rates could discourage HE participation especially among WP target groups.

Our recommendation is to postpone further consideration of employment metrics until the conclusion of the review of DLHE, and after conducting further research to assess the validity of employment outcomes as indicators of teaching quality and the suitability of alternative data sources. It is risky to introduce an under-researched metric which could have a negative impact on HE participation and confidence in the TEF. Furthermore, changes in the DLHE questionnaire are likely to lead to a metric that is not comparable with those for future years.

B) If included as a core metric, should we adopt employment in Standard Occupational Classification (SOC) groups 1-3 as a measure of graduates entering highly skilled jobs?

☐ Yes ☒ No ☐ Not sure

We reiterate the concern raised in response to A) that DLHE currently measures employment outcomes too early. In particular, it is unrealistic to expect employment in SOC group 1 (managers, directors and senior officials) within 6 months of graduation. Before a decision is made on the inclusion of a highly skilled employment metric, further research is necessary. Questions for investigation include: the length of time from graduation to employment in SOC groups 1-3, the amount of variation across HE providers in SOC 1-3 employment at 6, 12, 18 etc months after graduation, the factors associated with this variation, and the extent to which factors unrelated to teaching quality are adequately accounted for in the benchmarking procedures. SOC groups 1-3 would seem to capture most highly skilled jobs, although it is important to ensure any classification keeps pace with new growth areas in the labour market. We also share the concern mentioned in paragraph 71 of the consultation document that many jobs included in groups 1 and (particularly) 3 are not necessarily high value 'graduate jobs'.

C) Do you agree with our proposal to include all graduates in the calculation of the employment/destination metrics?

☐ Yes ☒ No ☐ Not sure

Please outline your reasons and suggest any alternatives.

The rationale given for the proposal to include all graduates in the denominator is to achieve consistency with internationally agreed definitions used in official employment measures, and to obtain a measure that shows more variation among providers than the current HESA indicator. However, the key question is which measure is most appropriate for evaluating the quality of HE providers.

There is substantial variation across providers in the proportion of DLHE respondents who are in the 'other activity' group (i.e. not working, studying or seeking work), ranging from 0 to 16.4% for UK domiciled leavers from full-time courses in 2013/14.[3] In contrast, the employment rate (excluding 'other activity' from the denominator) is relatively consistent across providers.[4] As reasons for not seeking work include taking time out for travel, illness and looking after home or family, it is likely that variation in the proportion in this group across institutions is due to differences in characteristics of the study body, rather than differences in teaching quality. Thus including this group in the denominator will distort comparisons of employment rates among providers. The impact of including 'other activity' will be greater for providers with fewer eligible graduates (e.g. high proportions of overseas students), especially when broken down by subject area. This is presumably the reason for HESA's decision to present an indicator that excludes 'other activity' from the denominator.

Question 3 (Chapter 3)
A) Do you agree with the proposed approach for setting benchmarks?

☐ Yes ☒ No ☐ Not sure

Our comments will be concerned with technical aspects of the proposed metrics and our assessment of what constitutes good practice in the reporting and use of results. Specifically we will consider issues of statistical bias and accuracy.

The National Student Survey (NSS)

Biases in estimates derived from surveys such as the NSS can arise in a number of ways. A moderate to large non-response rate may occur through non-random failure to respond and such biases may vary across individual institutions. Thus, the practice of 'gaming' almost certainly occurs, whereby some institutions incentivise a positive response to the survey. This practice is

[3] https://www.hesa.ac.uk/dox/performanceIndicators/1314_S5E7/e1a_1314.xlsx
[4] Based on 2013/14 figures for UK graduates from full-time courses, the coefficient of variation (100 × SD/mean) is 2.7 for the HESA indicator percentage working or studying (excluding 'other activity' from the denominator) compared with 47.6 for percentage 'other activity'.

almost always found in situations where 'high stakes' assessments are used and the further use of NSS within the TEF will lead to even higher stakes for institutions and their students. It is difficult to quantify such an effect but a process of quality control that attempts to monitor it is essential if credibility for the use of NSS is to be achieved. We note that the requirement of a (less than 100%) threshold response rate does not adequately address the bias issue.

Employment and further study post graduation

The issue of bias here is largely concerned with (non-random) failure to monitor all students. The response rate to the DLHE surveys is about 75%, but there appears to be little information about potential biases. Employment prospects differ by subject discipline and this clearly needs to be taken into account (benchmarked), and it appears that this is recognised already. Also, there are marked regional variations in employment opportunities and this will to some extent be related to University location, although taking account of it is not straightforward. Nevertheless, it is an issue that needs to be studied. We say more on this in our response to Q2.

Benchmarks

We are concerned that each institution is to be compared with an adjusted overall average (benchmark). Apart from a concern about the term 'benchmark' which usually applies a level having some externally agreed validity, the actual use of institutional ratings will often be to compare institutions one with another. It would seem inevitable that such comparisons will lead to 'league tables' and what is required, at the very minimum, is a way of presenting any such rankings in such a way, given the uncertainties, that misleading inferences are not drawn. This aspect needs to be addressed.

User understanding

There is undeniably a limited amount of public understanding of how to interpret results such as those proposed, especially in terms of possible bias and uncertainty based around small numbers. This will be particularly important for certain subjects where the numbers for any one institution will be small, even aggregated over a number of years. We therefore urge that attention is given to ways of presenting results so that the full nature of all the uncertainties of interpretation are preserved. The RSS has given attention to this in the past - for example in our work on 'Performance indicators: good, bad, and ugly', [5] and would be supportive of a similar endeavour in relation to the TEF.

B) Do you agree with the proposed approach for flagging significant differences between indicator and benchmark (where differences exceed 2 standard deviations and 2 percentage points)?

☐ Yes ☒ No ☐ Not sure

Please outline your reasons if you disagree.

[5] Bird, S., Cox, D., Farewell, V.T., Goldstein, H., Holt, T., Smith, P.C. (2005) 'Performance indicators: good, bad, and ugly' (PDF), *J. R. Statist. Soc. A*, *138*(1): pp. 1-27. Available from: http://www.rss.org.uk/Images/PDF/publications/rss-reports-performance-monitoring-public-services-2003.pdf

Uncertainty and significance

The proposal appears to be that a statistically significant difference from the 'benchmark' for any given institution will be judged at the conventional 95% level. It also appears that the metric will be in the form of a percentage and, that as well as being statistically significant, a substantive difference of 2 percentage points will be required for any institution to be 'flagged'. We agree with the principle of identifying differences that are both statistically and practically significant, but are concerned that going for this nominal level of significance may generate a substantial number of 'outliers', particularly if the results are over-dispersed. We note that document says "*the likelihood of a false flag (i.e. a difference exceeding the thresholds purely as a result of random variation) will increase - but to no more than 5% for any individual metric",* but this is the nominal chance that 'normal' institution gets flagged - it does *not* mean that only 5% of the signals are false positives. With many institutions, it is inevitable that many more of these signals will be false-positives. We note that other benchmarking schemes have adopted 3 standard deviations, and this more stringent criteria appears more appropriate and robust.

Question 4 (Chapter 3)

Do you agree that TEF metrics should be averaged over the most recent three years of available data?

☐ Yes ☐ No ☒ Not sure

Please outline your reasons and suggest alternatives.

We accept that some kind of aggregation over time is useful. Rather than a 3-year average, it would be better to present the current (latest) results together with an estimated (smoothed) trend over the most recent n years (n=3,4,5).

Question 5 (Chapter 3)

Do you agree the metrics should be split by the characteristics proposed above?

☐Yes ☐No ☒ Not sure

Please outline your reasons and suggest alternatives.

The benchmark is essentially the overall average within the 'split groups', such as gender, ethnicity etc. We believe that there is an important debate about which 'splits' should be included, but that is strictly outside our own remit and expertise, except to remark that institutions may have the ability to 'manipulate' their student body composition through e.g. acceptance criteria, in ways that the existing proposed 'splits' cannot capture, and that this may be important. We note that 'entry qualifications' is not to be included, yet this is predictive of success and presumably also associated with 'satisfaction'. It is possible to study the effect of such factors and there is an important debate about whether adjustment for such factors is needed. We are concerned that there is nothing about this in the consultation. Another way to present results is in terms of the probability that an institution with a score above the benchmark is in fact below it and vice versa. This is especially useful if benchmarks become used as quality thresholds.

It is not even clear that it is possible to discriminate the vast majority of HE institutions on the overall NSS satisfaction scores, let alone when they are broken down into smaller subgroups. Such a conclusion is supported by Figure 1 of "Teaching Excellence Framework: Review of Data Sources --- Interim Report" prepared by the Office for National Statistics' Methodology Advisory Service, February 2016 and the following text: "Confidence intervals are not available for breakdowns of NSS data by sub-groups such as ethnicity and socio-economic classification" and "comparisons of raw data between institutions at this level would not be statistically significant". [6] To extend such analyses down to the Department level is inappropriate.

Notwithstanding the lack of a category of `Does not meet expectations' (cf. Figure 9 and Question 12) what happens if benchmarking reveals to the TEF assessors that a provider "Meets Expectations" on one subgroup but does badly on another subgroup? This would be of interest to potential students in that latter group.

Question 6 (Chapter 3)
Do you agree with the contextual information that will be used to support TEF assessments proposed above?

☐Yes ☐No ☒ Not sure

Please outline your reasons and suggest any alternatives or additions.

It is important to provide assessors with some contextual information, and what is proposed will have some use. However, further clarity is needed on how the information will be provided. One issue is that the characteristics will not be independent of one another – for instance, entry qualifications are unlikely to be the same across different subject areas – and these relationships will be different in different institutions, making it difficult for the assessors to make appropriate use of the contextual information if it is presented too crudely.

Question 7 (Chapter 3)
A) Do you agree with the proposed approach for the provider submission?

☐Yes ☒No ☐ Not sure

Paragraph 101 on page 28 is troubling, and might have adverse consequences for the education of students studying STEM and quantitative skills subjects. [7] It is axiomatic that different disciplines often have quite different approaches to teaching and learning and asking assessors to "avoid

[6] P. 13 in ONS Methodology Advisory Service (2016) *Teaching Excellence Framework: Review of Data Sources - Interim Report* (PDF) Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/523291/bis-16-269-teaching-excellence-framework-review-of-data-sources-interim-report.pdf

[7] "Assessors will be looking for evidence of how far a provider demonstrates teaching and learning excellence across its entire provision. The submission should therefore avoid focusing on successful but highly localised practices that affect a relatively small number of students studying on particular courses or in particular departments", p. 101 in BIS (2016) *Teaching Excellence Framework Technical Consultation for Year Two* (PDF) Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/523340/bis-16-262-teaching-excellence-framework-techcon.pdf

focussing on successful but highly localised practices" seems to be against the spirit of the TEF. Good practice in any discipline is often localised and spread via discipline-specific peer groups (for example, learned bodies, professional institutions and associations) and less so across HE institutions. For example, many departments strongly prefer new university teachers to undertake national discipline-specific training (such as that provided in the past by the Maths, Stats and OR Network of the HEA) rather than generic courses provided by their HE institution.

B) Do you agree with the proposed 15 page limit?

☐Yes ☐No ☒ Not sure

Please explain your reasons and outline any alternative suggestions.

It seems clear that some limit is required although it is not clear precisely what the limit should be: 15 pages seems about right. However, the limit will inevitably have to grow in successive years as the TEF becomes more detailed and pervades deeper into subject disciplines.

Question 8 (Chapter 3)
Without the list becoming exhaustive or prescriptive, we are keen to ensure that the examples of additional evidence included in Figure 6 reflect a diversity of approaches to delivery. Do you agree with the examples?

☒Yes ☐No ☐ Not sure

Please outline your reasons and suggest any additions or alternatives?

All the examples are good in principle. Naturally, any list should not be prescriptive or exhaustive and good evidence should be welcomed and rewarded wherever and however it is found. Some of the examples possibly need a little more clarity; for example, the "Impact and effectiveness of external examining" could refer to the work of external examiners within the HEI in question, or the work of the HEI's staff in acting as external examiners elsewhere. In practice, it seems inevitable that institutions may provide additional evidence at a disciplinary level, rather than as averages across the entire institution. For example, some professionally-based disciplines will have extensive recognition of their courses by professional, statutory and regulatory bodies (PSRBs), while others will simply not have the opportunity to achieve this because no relevant PSRBs exist, and so a simple institution-wide listing of the recognitions may not be appropriate. Presenting additional evidence in a way that meets the requirements of paragraph 101 on submissions, to avoid focusing on localised practices, may therefore be somewhat challenging, and could thus lead to inequities in treatment of different institutions who set the balance between paragraph 101 and the call for additional evidence in different places.

Question 9 (Chapter 4)
A) Do you think the TEF should issue commendations?

☒Yes ☐No ☐ Not sure

B) If so, do you agree with the areas identified above?

☒Yes ☐No ☐ Not sure

Please indicate if you have any additional or alternative suggestions for areas that might be covered by commendations.

We think it could be useful to have a system of commendations to highlight and reward good practice and raise the profile of effective and innovative teachers. In particular, a commendation system is somewhat orthogonal to a metrics-driven statistical assessment that the rest of the TEF is intended to deliver. Commendations would presumably be awarded or assigned based on the professional judgement of TEF assessors and will not lead to `league tables'.

We do not see why providers could not receive unlimited commendations should the TEF assessors see fit, although clearly easily-won commendations will, over time, have less value. We are content that assessors would be given guidance not to be overly generous in their award of commendations, but are pleased to see (paragraph 121) that there is not an intention to force a distribution or quota.

An issue may be that, since the commendations are based on professional judgement, different assessors turn out to have different attitudes to issuing them, leading to inequity between assessments for different institutions. We hope that this can be mitigated by training and technical advice to the assessors, but recommend that this issue be kept under review.

We would be keen to see a class of commendations that were aligned with national strategic needs, such as education in modern languages, mathematical, quantitative or statistical subjects, where evidence to support such commendations has been presented, for instance as part of the additional evidence as exemplified in Figure 6. As we said in our reply to Question 8, such evidence is likely to be discipline-specific in some cases anyway.

Question 10 (Chapter 4)
Do you agree with the assessment process proposed?

☐Yes ☒No ☐ Not sure

Please outline your reasons and any alternative suggestions. The proposed process is set within a relatively tight timescale, reflected in the key dates included in Annex B. Responses should be framed within this context.

We agree with parts of it. It seems reasonable to have three stages, with the involvement of assessors, TEF officers, widening participation experts, employer representatives and panel members. A key question is what should the relative weighting of the metrics versus the additional evidence be? This is crucial and not an easy thing to determine. Given the expected small variability in metrics across institutions, the effective weight of the additional evidence could be very large. On the other hand, if the additional evidence was of high quality this could be a good thing, given the total inadequacy of the NSS for assessing teaching quality.

Question 11 (Chapter 4)
Do you agree that in the case of providers with less than three years of core metrics, the duration of the award should reflect the number of years of core metrics available?

☒Yes ☐No ☐ Not sure

Please outline your reasons.

Our experience of similar exercises, where data collection is required, is that --- although official data may amount to less than three years --- there will have been much ongoing work in advance of this (for example, Athena). So, we would not make this a blanket rule. However, the award duration could be modified by the panel if they are otherwise not confident of the sustainability of the application.

We do not see why appeals are excluded in Year Two (point 139). Providers who are dissatisfied with the process, and its outcome, will most effectively express this via an appeal and this can be incorporated into the review mentioned in that point. Further, there may be legal difficulties caused by the lack of the appeal route at this stage. The lack of appeals seems to go against the desire for the TEF to be transparent as mentioned in point 6.

Question 12 (Chapter 5)
Do you agree with the descriptions of the different TEF ratings proposed in Figure 9?

☐Yes ☒No ☐ Not sure

Please outline your reasons and any alternative suggestions.

It seems curious there is no descriptor that encapsulates unsatisfactory or poor performance. In particular, any provider entering the TEF that meets the “eligibility requirements” (point 15) will be guaranteed to receive at least a “Meets Expectations” rating. Such a system might cause considerable difficulties if a Year Two TEF assessment team subsequently find that a more detailed assessment (that they are carrying out) reveals substantial elements of poor performance. The TEF would then end up with a provider that “Meets Expectations” when the reality is much worse.

There is a disconnect between the metrics and the descriptors used to award `Excellent’ and `Outstanding’ compared to the key areas of the `Meets Expectations' which are about QAA quality and standards approval. We think it will be confusing unless the descriptions are made very clear.

Maybe the ratings should be:

Meets expectations in quality and standards
Meets expectations in quality and standards, and Excellent in three key areas
Meets expectations in quality and standards, and Outstanding in three key areas.

Otherwise, we think it is likely that the higher TEF ratings will erroneously be interpreted as:

`Excellent in quality and standards’ or `Outstanding in quality and standards’

where `quality and standards' are not being properly measured by the metrics or the panel.

[Response ends]

Response submitted by RSS' Policy and Research Manager, 11 July 2016