ROYAL STATISTICAL SOCIETY:

RESPONSE TO THE TEACHING EXCELLENCE AND STUDENT OUTCOMES FRAMEWORK, SUBJECT-LEVEL CONSULTATION

The Royal Statistical Society (RSS) was alarmed by the serious and numerous flaws in the last Teaching Excellence Framework (TEF) consultation process, conducted in 2016. Our concerns appeared not to be adequately addressed by the Department for Education (DfE). Indeed, the DfE's latest TEF consultation exercise, which will shortly close, suggests that few statistical lessons have been learned from 2016's experience. As we argue, below, there is a real risk that the latest consultation's statistically inadequate approach will lead to distorted results, misleading rankings and a system which lacks validity and is unnecessarily vulnerable to being 'gamed'.

Background: 2016's TEF consultation

1. In July 2016, the Royal Statistical Society (RSS) responded to the Government consultation exercise on the *Teaching Excellence Framework: Year Two and Beyond*.[i]

2. The RSS's response focused almost entirely on the important statistical and scientific shortcomings that we identified in this consultation paper.[ii]

Many aspects of the proposed methodology caused us serious unease. Our main concerns stemmed from:

a. the lack of an adequate statistical underpinning for many of the processes described in the consultation paper
b. the ways in which various employment metrics had been used
c. the paper's assumptions around causality
d. the lack of evidence about a link between teaching quality and employment outcomes
e. the use of DLHE-like metrics that distort employment rate comparisons between employers due to student demographic factors *(DLHE: Destinations of Leavers from Higher Education)*
f. non-response levels in the National Student Survey (NSS)
g. 'gaming' in metrics, particularly in the case of the NSS
h. the clearly inadequate way in which uncertainty had been handled
i. the weighting of metrics and how variability affects this
j. problems with the medal-like Teaching Excellence Framework rankings and the assessment of their uncertainty.

3. Some of our concerns were shared elsewhere. The Office for Statistics Regulation (OSR) wrote to the Department for Education on 23rd February 2017 and asked it to "Ensure that the … concerns raised by the RSS in their TEF consultation response in July 2016 have been addressed and published." Although the DfE had published a response in September 2016, we believe it was inadequate and we did not sense that our substantive concerns had been addressed.[iii] [iv]

4. We were particularly concerned as the DfE appears to be making scientific decisions by unscientific polling of respondents, rather than via scientifically credible methods. Such conduct

should be the cause of serious concern among all those who share our belief in the importance of statistics, science and rigour within government decision-making.

5. In 2016’s consultation, for example, Question 5 on “TEF Metric Splits” suggested that the DfE had made a decision based on “the majority of respondents” rather than assessing whether the decision made statistical sense, or whether one could trust the numbers, especially when so little was known about the underlying uncertainty.

6. We note that the Office for National Statistics’ report *Teaching Excellence Framework* has also commented, at Section 3.2.3, on the statistically non-informative nature of these kinds of findings.[v]

7. As 2016’s experience showed, we do not believe that such a consultation process represents an acceptable way of guiding the statistical design of an assessment framework. Many of the questions posed by the consultation paper were technical statistical ones, which can only be properly addressed and answered through professional statistical assessment - not by unscientific polls.

8. We were (and remain) concerned about decisions on assessment design being made by polling the likely future subjects of the assessments themselves. This does not appear to represent a good route to obtaining a rigorous and unbiased assessment - especially with, in the case of 2016’s consultation, 75% of the responses coming from education providers or students’ unions. We believe that such biases could have significant, serious and unintended consequences in the years ahead.

9. We will turn, next, to the current Subject Level consultation being conducted by the DfE. Again, we believe that a consultation exercise of this sort represents an inappropriate means of answering many of the key statistical questions that are being posed. Decisions should be based, instead, on appropriate statistical and scientific evaluation. Accordingly, the RSS considers that the current TEF process of evaluation and ranking is not statistically robust: as a result, the process lacks validity and might even leave the DfE open to legal challenge.

2018’s subject-level consultation

10. The RSS has numerous concerns about the current TEF consultation. The following list is not exhaustive, but summarises our main concerns:

a. Sections 7 and 8, p. 17. Model A “by exception” seems to be designed primarily to reduce costs, but at the expense of inducing bias. Our understanding of its operation is that: (i) a provider-level rating is produced for each provider; (ii) subject-level metrics are computed for each subject and provider; and (iii) where the subject-level metric is “different from” the provider rating, the subject is then assessed in more detail to result, potentially, in a new rating.

b. As previously mentioned in respect of our 2016 response, we are concerned that provider-level ratings are themselves flawed. However, putting that temporarily aside, there are additional problems with Model A. As an illustrative example, suppose we have a simplified system in which all the subject-level ratings across all institutions have the same statistical distribution. Then, just by chance, some institutions will receive overall low ratings and some will receive high ones. Subjects in institutions whose overall rating was low will be much more likely to be determined as ‘exceptions’ by this method and, consequently, when studied in more detail, will result in ratings higher than their provider's overall assessment. Hence, ironically, subjects with a low overall score will possess proportionately more subjects with higher ratings their institution. A similar, but opposite argument, applies for those institutions whose ranking was high overall, due to chance. In this situation, of course, we know that the distribution across all subjects in this illustrative example is the same. So, even when there are no real differences between subjects and institutions Model A, would tend to produce misleading rankings.

c. Question 6, p. 18. The consultation paper asks, "In Model A, should the subject ratings influence the provider rating?" On its own, this seems to be a curious thing to wish to do, as the subject ratings are only generated conditional on their 'distance' from the original provider rating. Furthermore, the consultation document does not explicitly consider the time series nature of the various metrics and rankings. For example, how do previous rankings from the National Student Survey (NSS) - or other metrics, for that matter - influence students who are currently completing NSS questionnaires? Question 6 suggests introducing (perversely, in our view) a new, fairly arbitrary source of feedback into the system, which could distort future results and make such systems harder to study. Moreover, there appears to be very little evidence for what sampling rate (period before reassessment) is appropriate in this context.

d. In our response to 2016's consultation, we noted that higher education was subject to both the Research Excellence Framework (REF) and the TEF, with the separate assessments failing to take proper account of the vital interplay between research and teaching in many institutions. Although the two assessments consider similar numbers of subject areas (34 for REF; 35 for TEF), it is striking that subjects are grouped in different ways for the two assessments: there are four main panels for the REF and seven subject groups for the TEF. Moreover, these groupings contain different subjects. For example, agriculture falls into "Medicine, health and life sciences" in the REF but "Natural science" in the TEF.

e. It is also noticeable that the number of subjects in each Group (as shown in Table 4, p. 36) is highly variable, ranging from just one subject in a group (the Arts group, which contains creative arts & design) to eight subjects (in the Humanities group). We believe that the nature and style of the different subject group submissions in Model B will vary widely from each other - not necessarily due to disciplinary differences, but also due to the widely varying number of subjects in each group.

f. The grouping proposal seems even more problematic when one realises that different institutions often have (i) quite different mixtures of subjects and (ii) different 'locations' of subjects. For example, many, but not all, institutions have a Department of Mathematics, but some subjects (e.g. agriculture, architecture and parts of medicine, etc) are present in fewer institutions. In the case of (ii), an example we know well, there are Departments of Mathematics in Schools/Faculties of Science, Social Science, Computing and various branches of Engineering. Both (i) and (ii) mean that the subject group submissions from different institutions simply will not be comparable and surely invalidate fair assessment? The consultation itself has partially recognised this problem (see Section 7, Subject Groups, p. 10) but its proposed solution is to permit providers to move one subject in and out of each group. In the RSS's assessment, this would actually worsen the comparability problem.

g. Section 6.2, p. 16. In Model B, it is not transparent how the subject metrics would be combined with the subject group submission to result in individual subject ratings. A similar operation occurs in the case of the REF, where panels have the latitude to weigh output and impact rankings, contextual information and metrics from environmental submissions to come to a final profile (which is more appropriate, in our assessment, than TEF's coarse bronze, silver and gold ratings). However, the REF sub-panels cover their, and only their, discipline. In the TEF Model B system, the subject group submission would presumably be considered by the subject group panel and hence influence several subjects simultaneously. For example, the teaching approach to agriculture could influence the assessment of either the mathematical sciences or geography (which might require extensive fieldwork), or generally unduly influence the assessment discussions. These cross-disciplinary pressures on assessment do not happen to the same extent in the REF and the RSS is not keen on their presence here. Put simply, the current suggested approach tries to assess Subject X but would be influenced by Subject Y. The subject grouping structure behind Model B will probably save money, but we are concerned that it will have a negative impact on the overall fairness of the assessment.

h. Core metrics, p. 24. The process for addressing non-reportable core metrics (see Figure 8) is arbitrary and highly subject-specific.

i. Equality and diversity, p. 14. The consultation has paid only token attention to equality and diversity issues. The sole mention seems to be within the last bullet point in the Question 4 box. Arguably, overall improvements in equality and diversity might be more effectively driven by improvements in specific subjects where, in some cases, there are apparent inequalities and a lack of diversity.

j. Section 13, p. 28. A significant proportion of the consultation is devoted to teaching intensity measures. The consultation document rightly identifies that there are many general principles that need to be taken into consideration and that "any teaching intensity measure should capture this diversity". Unfortunately, the direction of travel seems to involve the use of simple intensity measures - mostly capturing teaching time and the numbers of students per class. The reasons behind this approach being adopted seem to include the simplification of data capture and, hence, the reduction of costs. As a result, however, the RSS believes that the measures appear too simplistic and too liable to be being 'gamed', without contributing to improved teaching in higher education. We think it would be premature (to say the least) to include such metrics in the assessment and believe that more, and higher quality, research should be carried out to assess the genuine efficacy of such measures.

k. Pages 20 and 22. The RSS is unclear why it is proposed to use the same methods (using existing factors and groupings) to benchmark subjects as in the institutional-level assessment. If it is believed that benchmarking is a process to group similar universities, then subject-level benchmarks have to be different in different subjects. There appears to be a severe logical inconsistency here.

l. Page 21. As the consultation document appears to believe that subject-specific ratings have distributions which vary naturally, this would seem to cast doubt on the validity of the TEF - certainly the institutional-level TEF and particularly in the cases of institutions with a heterogeneous subject mix.

m. Section 10.3, p. 23, Distribution of Subject Rankings. The consultation document makes a compelling case for the difference in subject-level ratings - not just in terms of overall scale, but also as the ratings for some institutions exhibit different levels of clustering. In statistical science, such problems often arise and it is not clear that the 'do nothing' proposal is the correct approach - as quantities are being compared that naturally arise on different scales. One might propose a form of re-scaling, but the institution/subject-specific clustering would seem to cause severe problems with this, too. Essentially, the consultation paper identifies a problem with the data that its analysis cannot overcome; the result would almost certainly be flawed assessment.

n. These issues could cause even more severe problems when one considers joint and multi-subject programmes. For example, suppose you have three disciplines, whose ratings are on different scales, which may include clustering and properly subject-benchmarked in different ways for different subjects for different institutions. We do not yet understand how these data are to be combined so as to obtain, or contribute to, a robust and valid ranking.

o. Section 13.2, *p. 29*. Overall, the paper appears to give inadequate recognition to ‘gaming’ or the associated Goodhart principle. This seems intellectually dishonest. The problem is mentioned indirectly in this section: “Furthermore, the Government considers it important that data collection in this area should not itself drive teaching practices, nor impinge institutional autonomy by mandating activities that a provider may consider unfavourable to students or contradictory to its ethos of teaching, such as mandatory attendance monitoring” - but there is nothing to suggest how ‘gaming’ would actually be prevented.

p. More generally, there appears to be little evidence that students are currently making poor choices: furthermore, there is no evidence that they would make better choices in the future if they were additionally provided with poor quality subject assessments. There also appears to be no cost-benefit analysis to assess whether the effect of adding very high-quality subject assessment (not the current TEF) to the performance of UK plc would outweigh the increased cost of the assessment process.

Section 5, p. 11. As with school tables (see Leckie and Goldstein (2009), *Journal of the Royal Statistical Society A*, 172, 835-851), estimation of correlations across time permit one to assess the degree to which results can inform new students about the likely outcome of their final degrees. For TEF as a whole, we do not know what these correlations are and, again, this casts doubt on the validity of the exercise. This feature / problem is built-in to any timeframe for reassessment, but the problem is exacerbated when extending the duration between re-applications

[i] https://www.gov.uk/government/consultations/teaching-excellence-framework-year-2-technical-consultation

[ii] http://www.rss.org.uk/RSS/Influencing_Change/Higher_education/Higher_education_policy/RSS/Influencing_Change/Higher_education_policy/Higher_education_policy.aspx?hkey=ac4d01b6-8d3e-4e41-89b3-15b68e9b3de3

[iii] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/557140/Teaching_Excellence_Framework_-_Technical_Con_Response.pdf

[iv] http://www.rss.org.uk/RSS/Influencing_Change/Higher_education/Higher_education_policy/RSS/Influencing_Change/Higher_education_policy/Higher_education_policy.aspx?hkey=ac4d01b6-8d3e-4e41-89b3-15b68e9b3de3

[v] 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

Royal Statistical Society
21st May 2018